Detectando Doença de Parkinson - Uma Comparação de Modelos de Aprendizagem de Máquina com Redução de Dimensionalidade Diferencialmente Privada

Manuel E. B. Filho[1], **Maria de Lourdes M. Silva**[1], **Patricia V. da S. Barros**[1], **César L. C. Mattos**[1]

[1]Departamento de Computação (DC) – Universidade Federal do Ceará (UFC)
CEP 60440-900 – Fortaleza – CE – Brazil

{edvar.filho, malu.maia, patricia.barros}@lsbd.ufc.br, cesarlincoln@dc.ufc.br

Abstract. *This paper aims to present a comparison of machine learning models using two dimensionality reduction approaches in data pre-processing, one private and one non-private, to classify patients as having or not Parkinson's disease. After the data pre-processing step, seven machine learning models are used and compared with each other based on the analysis of metrics that seek to verify the models' behavior in diagnosing Parkinson's disease from a vocal collection. With the analysis of the results, the Gaussian process and the Random Forest, are the models that obtained the best performance in approach without privacy and with privacy, respectively, and new investigations are proposed.*

Resumo. *O presente artigo visa apresentar uma comparação de modelos de aprendizagem de máquina utilizando duas abordagens de redução de dimensionalidade no pré-processamento dos dados, uma privada e outra não privada, para classificar pacientes como portadores ou não da doença de Parkinson. Após a etapa de pré-processamento dos dados, sete modelos de aprendizagem de máquina são usados e comparados entre si a partir da análise de métricas que buscam verificar o comportamento dos modelos em diagnosticar a doença de Parkinson a partir de uma coleta vocal. Com a análise dos resultados, o processo Gaussiano e o Random Forest, são os modelos que obtiveram melhor performance em abordagem sem privacidade e com privacidade, respectivamente, e novas investigações são propostas.*

1. Introdução

A doença de Parkinson é um distúrbio neurológico causado pela degeneração de uma pequena parte do cérebro, afetando o controle de movimentos de indivíduos e coordenação motora, atualmente a doença de Parkinson é a segunda doença neurodegenerativa mais comum e geralmente acomete pessoas com idade superior a 60 anos, podendo causar agitação involuntária de membros, lentidão ou ausência de movimento, inflexibilidade das articulações e perda de equilíbrio.

Segundo estudos, cerca de 90% das pessoas diagnosticadas com doença de Parkinson exibem algum tipo de distúrbio vocal na fase inicial da doença, por isso a maioria dos estudos recentes que visam diagnosticar este distúrbio concentram-se na detecção de deficiências vocais [Erdogdu Sakar et al. 2017, Sakar and Kursun 2010].

Com a finalidade de auxiliar no avanço dos estudos para a realização do diagnóstico da doença com maior índice de veracidade, os pacientes, saudáveis ou não, devem abrir mão de seus dados de saúde para que este processo possa ser realizado com êxito. No entanto, a divulgação de dados de saúde deve ser feita garantindo a privacidade dos pacientes cujos registros serão publicados, de forma que não inutilize os dados para que possam ser feitas análises.

Dados médicos sobre as condições dos pacientes costumam implicar na qualidade dos tratamentos. Quanto mais dados do paciente, mais informações para guiar o profissional de saúde na tomada de decisão médica. Esses registros eletrônicos de saúde costumavam ficar sob responsabilidade da instituição no qual o paciente havia sido atendido, entretanto hoje esses registros podem ser disponibilizados através de varias instituições de saúde [Haas et al. 2011]. Existem empresas que podem gerir esses dados, essas empresas podem utilizar e vender informações obtidas através desses dados para outras instituições, devendo respeitar as leis que regem o armazenamento e o uso desses dados sensíveis de indivíduos cujos dados foram disponibilizados.

Visto a importância da privacidade dos indivíiduos cujos dados de saúde foram publicados, o presente artigo possui o objetivo de elaborar uma metodologia capaz de classificar se um indivíduo possui ou não a doença de Parkinson, além de comparar os resultados dos modelos de aprendizagem de máquina aplicados aos dados privados e dados originais dos pacientes.

No decorrer do artigo, será feita a análise de trabalhos relacionados, tal como a descrição da forma metodológica adotada para a avaliação dos modelos considerados. Os experimentos computacionais são feitos a partir de um conjunto de dados extraídos do Departamento de Neurologia da Faculdade de Medicina da Universidade de Istambul, disponibilizados na plataforma Kaggle. No final, os resultados dos diferentes modelos são discutidos e direções para investigações futuras são apontadas.

2. Trabalhos Relacionados

Pesquisas sobre detecção de doença de Parkinson a partir de análise de coletas vocais já é bastante consolidado, mas de forma a garantir a privacidade dos pacientes testados ainda é algo em estágio inicial, já que é algo relativamente recente, pelo menos no que diz respeito ao interesse levantado pela sociedade. Analisamos três (3) trabalhos que possuem o foco similar ao objetivo aqui apresentado, em relação a análise de modelos de aprendizagem de máquina e algum tipo de redução de dimensionalidade ou seleção de atributos:

Em [Ramani and Sivagami 2011] foi realizado a busca da melhor regra de classificação para diagnosticar de maneira correta pacientes saudáveis e com doença de Parkinson a partir da análise comparativa entre regressão logística, máquinas de vetores de suporte (SVM), *Random Forest*, análise de discrimante linear, entre outros. A comparação dos modelos entre si é feita a partir da análise da matriz de confusão e das taxas de erros apresentadas por cada modelo, que foram treinados a partir de atributos considerados relevantes para o problema, selecionados a partir da relevância do atributo para o modelo.

Em [Das 2010] é feito a comparação entre a regressão logística, árvore de decisão, rede neural, e *DMNeural*, mas sem trazer uma abordagem que garanta a privacidade dos usuários, tem uma ótima acurácia em classificar os pacientes, que é resultante do modelo

de rede neural, que se sobressai aos outros comparado-os a partir da curva ROC. Os resultados dos modelos deste trabalho são obtidos a partir de uma seleção de variáveis de entrada, que visa selecionar o melhor conjunto de entrada, para que se obtenha um resultado positivo.

Os dois trabalhos apresentados anteriormente, utilizam o mesmo conjunto de dados criado pela Universidade de Oxford, e contém um total de 195 registros vocais, com 23 atributos, de um total de 31 pessoas, ou seja, há mais de uma coleta vocal atribuída a mesma pessoa, o que acarreta resultados tendenciosos, pois, pode haver registros de um mesmo usuário tanto no conjunto de treinamento, como no conjunto de teste.

Em [Sakar et al. 2019] é criado um conjunto de dados diferentemente do anterior, que utilizamos em nosso trabalho, resultado de três coletas atribuídas a cada um dos 252 pacientes que realizaram o exame vocal, dentre os quais 188 possuem a doença de Parkinson, e há um total 755 atributos gerados a partir de diversas coletas e transformações e que formam conjuntos de atributos. O trabalho realiza um coparativo entre regressão logística, K-*Nearest Neighbors* (KNN), *Random Forest*, e diversos outros modelos de aprendizagem, que foram comparados analisando a acurácia, *F1-Score* e o coeficiente de correlação de Matthews.

Com base nas análises feitas anteriormente, o nosso trabalho busca aderir práticas presentes nos trabalhos citados, desenvolvendo, assim, uma metodologia para classificação de doença de Parkinson, que inicia no pré-processamento dos dados, com o uso de técnicas de redução de dimensionalidade em abordagens com privacidade e sem privacidade por meio de combinação de atributos, até a comparação de modelos de aprendizagem de máquina, cujos hiperparâmetros são selecionados por meio da técnica *grid search*, com a devida separação em conjuntos de treino, validação e teste, e o uso de validação cruzada por meio de *k-folds* para um valor de k igual a 5.

3. Metodologia

O problema considerado no presente artigo trata-se de uma classificação binária, em que um paciente pode ser rotulado como *portador* ou *não portador* da doença de Parkinson. O conjunto de dados utilizado foi extraído do Kaggle [Kaggle 2020] e consiste em dados coletados de 252 pacientes, dentre os quais 188 possuem doença de Parkinson, 107 homens e 81 mulheres com idades variando entre 33 e 87 anos, e 64 indivíduos saudáveis, 23 homens e 41 mulheres com faixa etária entre 41 e 82 anos. Cada paciente possui três registros referentes ao seu quadro.

Os atributos originais do conjunto de dados foram obtidos a partir de vários algoritmos de processamento de sinais de voz aplicados às gravações de fala dos pacientes submetidos a coleta, incluindo frequência temporal, coeficientes cepstrais de frequência Mel, atributos baseados em transformação Wavelet, atributos de dobra vocal e atributos resultantes da transformada Wavelet de fator Q ajustável, além dos atributos básicos, como o identificador do paciente, gênero, idade, e outros relevantes para o problema, além da classificação do paciente, como portador ou não da doença, totalizando assim 756 atributos [Sakar et al. 2019].

O procedimento de pré-processamento dos dados e redução de dimensionalidade, treinamento e avaliação dos modelos será descrito nas seções abaixo.

Os códigos em Python usados para gerar os resultados dos experimentos podem ser encontrados no repositório `https://github.com/EdvarFilho/TrabalhoAprendizagemAutomatica`.

3.1. Privacidade Diferencial e Mecanismo Gaussiano

A privacidade diferencial foi um modelo proposto pela matemática Cynthia Dwork [Dwork et al. 2014], para garantir a privacidade de indivíduos em um conjunto de dados.
Definição 1 (Privacidade Diferencial)**.** Um mecanismo *M* é (ϵ, δ)-diferencialmente privado se:

i para todos os *datasets* vizinhos D_1 e D_2, onde *datasets* vizinhos são conjuntos de dados que se diferem em apenas um registro;
ii para todo conjunto *S* contido na variação de resultados de *M*, isto é, para todo $S \subset Range(M)$.

A seguinte condição deve ser satisfeita:

$$Pr[M(D_1) \in S] \leq \exp(\varepsilon) \times Pr[M(D_2) \in S] + \delta. \tag{1}$$

Sendo ε, um limite para a perda de privacidade de uma consulta e δ a relaxação da equação.

Neste trabalho, utilizaremos o mecanismo Gaussiano [Dwork et al. 2006] que adiciona ruído Gaussiano com média 0 e variância σ^2 na saída de uma consulta, para torná-la privada.
Definição 2 (Sensibilidade l_2)**.** Para uma consulta $f : D \to \mathbb{R}$, a sensibilidade l_2 da função f é definida como:

$$\Delta_2 f = \max \|f(D) - f(D')\|_2. \tag{2}$$

Definição 3 (Mecanismo Gaussiano)**.** Dada uma função $f : D \to \mathbb{R}$ sobre um *dataset* D, se $\sigma = \Delta_2 f \sqrt{\frac{2 \times \ln(\frac{2}{\delta})}{\epsilon}}$ e $N(0, \delta^2)$ uma variável randômica Gaussiana independente e identicamente distribuída, o mecanismo M provê (ϵ, δ)-privacidade diferencial, quando segue:

$$M(D) = f(D) + N(0, \delta^2) \tag{3}$$

3.2. Pré-processamento e Redução de Dimensionalidade

Devido a grande quantidade de atributos dispostos ao problema, a execução de alguns modelos seriam diretamente afetados no que diz respeito ao tempo de processamento e execução. Além disso, os dados possuem escalas diferentes, com isso como etapa de processamento foi realizada uma normalização dos dados, com exceção do identificador do paciente, que não influencia o resultado do processo de classificação, e a classe do paciente.

Para reduzir a dimensão dos dados, trouxemos uma abordagem privada e outra não privada. Com a abordagem não privada aplicamos a técnica de Análise de Componentes Principais (*Principal Component Analysis* - PCA), capaz de combinar linearmente os atributos em um espaço de menor dimensão. Buscamos assim obter os dados com dimensão

de tamanho 100, obtidos a partir da seleção dos 100 autovetores relacionados aos 100 autovalores calculados a partir da decomposição de valor singular da matriz normalizada dos dados.

Visando observar o comportamento dos dados a partir de uma redução de dimensionalidade diferencialmente privadas, aplicamos o algoritmo inicial Mod_SULQ prospoto em [Chaudhuri et al. 2012], que consiste em uma modificação realizada no algoritmo SULQ apresentado em [Blum et al. 2005], este que gerava uma matriz não simétrica para obter os autovalores e autovalores referentes a ela, podendo assim gerar autovetores complexos.

O algoritmo Mod_SULQ que garante (ϵ, δ)-privacidade diferencial, realiza uma transformação na matriz normalizada dos dados, tornando-a simétrica, e por meio dos dados dimensionais da matriz original normalizada dos dados e com os parâmetros de privacidade ϵ e δ, configurados como 1 para ambos, é calculado o desvio padrão de um distribuição Gaussiana com média 0. A partir dessa distribuição, geramos uma matriz de ruído com as dimensões da matriz quadrada gerada, que é adicionada a esta, e todos os valores da matriz de ruído seguem essa distribuição Gaussiana de maneira independente e identicamente distribuídas. Com essa matriz quadrada resultante, obtemos desta os 100 maiores autovalores e os autovetores referentes a cada um, para termos a matriz de transformação que diminui a dimensão dos dados normalizados.

3.3. Modelos de Aprendizagem Supervisionada

Após a limpeza dos dados, separamos os conjuntos de treino e teste, com uma porcentagem de 25% dos dados para teste, de modo que todas as coletas de um mesmo paciente ou estarão no conjunto de treinamento ou no conjunto de teste, para que não haja dados do mesmo paciente em ambos os conjuntos e os resultados sejam tendenciosos. Os modelos de aprendizagem supervisionada utilizados foram: Regressão Logística (RL), Análise do Discriminante Gaussiano (AGD), K-*Nearest Neighbors* (KNN), Árvore de Decisão (AD), *Random Forest* (RF), Máquina de Vetores de Suporte (SVM) e Processo Gaussiano (PG). Os algoritmos em questão foram selecionados por serem técnicas tradicionais de aprendizagem de máquina amplamente usados na prática em tarefas de classificação.

Os hiperparâmetros dos modelos, quando necessário, foram selecionados via *grid search*, onde foi aplicada a função *GridSearchCV* disponibilizada pela biblioteca *scikit-learn* [Pedregosa et al. 2011] para todos os modelos, com exceção da Regressão Logística cuja seleção de hiperparâmetros foi implementada manualmente. Tal função realiza a divisão do conjunto de treino em treino e validação, e realiza uma validação cruzada por meio da técnica *k-fold* usando o valor de *k* igual a cinco (5), exceto para a regressão logística em que 25% dos dados disponíveis para treino, são utilizados como conjunto de validação para a seleção dos hiperparâmetros.

3.4. Métricas de Avaliação

As métricas utilizadas para avaliar os modelos são: acurácia, que mede o quão próximo estão as classes reais das classes preditas pelo modelo; precisão, que verifica se, dados os exemplos classificados como verdadeiros, quais realmente são verdadeiros, dando assim uma fração de instâncias recuperadas que são relevantes; revocação, que dá a fração de instâncias relevantes que são recuperadas; *F1-Score*, que é utilizado em problemas de

classificação binária sendo uma média harmônica entre a precisão e a revocação; e a curva ROC (*Receiver Operating Characteristic*) e consequentemente a área sobre a curva ROC, que ilustra o desempenho de um classificador a partir das taxas de falsos positivos e verdadeiros positivos, e a análise da curva, respectivamente.

A comparação dos resultados obtidos será feita ainda através de matrizes de confusão, ou seja, tabelas que mostram as frequências de classificação para cada classe do modelo, indicando os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

4. Experimentos

Em nossos experimentos, realizamos uma divisão do conjunto de dados para treinamento e para teste. Para treino temos um total de 567 (75% dos dados) registros e para teste o valor de 189 (25% dos dados) registros, como já citado anteriormente.

Em seguida, realizamos o procedimento *grid search*, fazendo uso da função *GridSearchCV*, citada anteriormente, que também realiza uma validação cruzada por meio da técnica *k-fold* usando o valor de *k* igual a 5 e a realiza a divisão dos dados em validação e treino, para escolher os melhores hiperparâmetros para os modelos propostos, com exceção da Análise de Discriminante Gaussiano, que não possui hiperparâmetros e o modelo de Regressão Logística, que implementamos a função de seleção dos hiperparâmetros. Para cada um dos algoritmos de aprendizagem de máquina de classificação temos os hiperparâmetros testados e os melhores resultados obtidos após a execução do *grid search* estarão presentes logo abaixo.

- **Regressão Logística:** Hiperparâmetro: Quantidade de épocas, tamanho do passo de aprendizagem (α), fator de regularização (λ).
  - **Valores testados:** Quantidade de épocas: 100, 1000, 10000. Valores de α: 0.001, 0.01, 0.1, 1. Valores de λ: 0.01, 0.1, 1.
  - **Valores escolhidos na abordagem não privada:** Quantidade de épocas: 1000. α: 0.1. λ: 0.01.
  - **Valores escolhidos na abordagem privada:** Quantidade de épocas: 10000. α: 0.1. λ: 0.01.
- **Árvore de Decisão:** Hiperparâmetros: Profundidade máxima da árvore e índice de pureza.
  - **Valores testados:** Profundidade máxima: 1 a 50, variando em apenas uma unidade. Índice de pureza: entropia ou índice de Gini.
  - **Valores escolhidos na abordagem não privada:** Profundidade máxima: 1. Índice de pureza: entropia.
  - **Valores escolhidos na abordagem privada:** Profundidade máxima: 5. Índice de pureza: Gini.
- **SVM:** Hiperparâmetros: Função de kernel, valor de margem (C), gama dependendo da função de kernel, assim como o grau do polinômio.
  - **Valores testados:** Função de kernel: Gaussiano ou polinomial. Valor de C: 2 com potência de -3 a 2 variando-as em uma unidade. Valor de gama para kernel Gaussiano: 2 com potência de -3, -2 e -1. Grau do polinômio para kernel polinomial: 3, 4 a 5.

  - **Valores escolhidos na abordagem não privada:** Função de kernel: polinomial. Valor de C: 0.125. Grau do polinômio: 3.
  - **Valores escolhidos na abordagem privada:** Função de kernel: polinomial. Valor de C: 4. Grau do polinômio: 3.
- ***Random Forest*:** Hiperparâmetros: Número de estimadores, profundidade máxima para as árvores e índice de pureza.
  - **Valores testados:** Número de estimadores: 100 a 150, variando em uma unidade. Profundidade máxima: 1 a 50, variando em apenas uma unidade. Índice de pureza: entropia ou índice de Gini.
  - **Valores escolhidos na abordagem não privada:** Número de estimadores: 146. Profundidade máxima: 9. Índice de pureza: índice de Gini.
  - **Valores escolhidos na abordagem privada:** Número de estimadores: 137. Profundidade máxima: 8. Índice de pureza: índice de Gini.
- **KNN:** Hiperparâmetro: Número de vizinhos (K).
  - **Valores testados:** Valor de K: 3 a 11, variando em duas unidade.
  - **Valor escolhido na abordagem não privada:** Valor de K: 9.
  - **Valor escolhido na abordagem privada:** Valor de K: 5.
- **Processo Gaussiano:** Hiperparâmetro: Função de kernel.
  - **Valores testados:** Função de kernel: Gaussiano, Materno, Quadrático Racional, Periódico.
  - **Valor escolhido na abordagem não privada:** Função de kernel: Quadrático Racional.
  - **Valor escolhido na abordagem privada:** Função de kernel: Materno.

5. Resultados

Após o treinamento com os melhores hiperparâmetros selecionados anteriormente, que divergem entre as abordagens privada e não privada devido ao ruído adicionado à matriz dos dados para a redução de dimensionalidade, foram realizados testes com 25% dos dados originais e os resultados obtidos para a abordagem sem privacidade e com privacidade estão apresentados nas Tabelas 1 e 2, respectivamente. Pode-se perceber que, com relação a todas as métricas avaliadas, os modelos em ambas abordagens possuem resultados semelhantes, onde a variação dos resultados, independente do modelo, ocorre devido a adição do ruído Gaussiano.

Tratando-se da abordagem sem privacidade, a aplicação do Processo Gaussiano para o problema de classificação de doença de Parkinson, obtém resultados positivos e convincentes, apresentando a melhor acurácia, precisão, revocação e *F1-Score*, com os valores de 81%, 82%, 81% e 78%, respectivamente. O processo Gaussiano, obtém esse resultado positivo, pois ele é um modelo probabilístico, embasado em interpoladores de dados, que por meio de uma função de kernel consegue realizar um mapeamento dos dados, utilizando essa função como função de distância entre os dados e que vai definir a matriz de covariância de uma distribuição normal multivariada com média 0.

Na abordagem privada, o modelo que obteve melhor resultado foi o *Random Forest*, com acurácia, precisão, revocação e *F1-Score* sendo respectivamente de, 81%, 81%, 81% e 79%, que por se tratar-se de um comitê de árvores de decisão, é possível observar

Tabela 1. Valores das métricas - Abordagem não privada

Modelo	Acurácia	Precisão	Revocação	F1-Score	AUC
Regressão Logística	0.68	0.76	0.68	0.70	0.79
Análise de Discrimante Gaussiano	0.76	0.73	0.76	0.71	0.64
Árvore de Decisão	0.76	0.73	0.76	0.74	0.62
SVM	0.80	0.78	0.80	0.78	0.71
Random Forest	0.80	0.80	0.80	0.76	0.75
KNN	0.79	0.78	0.79	0.76	0.63
Processo Gaussiano	**0.81**	**0.82**	**0.81**	**0.78**	**0.83**

Tabela 2. Valores das métricas - Abordagem privada

Modelo	Acurácia	Precisão	Revocação	F1-Score	AUC
Regressão Logística	0.67	0.75	0.67	0.69	0.76
Análise de Discrimante Gaussiano	0.75	0.71	0.75	0.67	0.62
Árvore de Decisão	0.75	0.73	0.75	0.74	0.70
SVM	0.79	0.79	0.79	0.75	0.70
Random Forest	**0.81**	**0.81**	**0.81**	**0.79**	**0.81**
KNN	0.79	0.77	0.79	0.75	0.62
Processo Gaussiano	0.79	0.78	0.79	0.75	0.81

que a aplicação da árvore de decisão já traz um resultado favorável e com a quantidade de estimadores que temos, possuímos um melhor resultado, apesar de que devido a quantidade de estimadores selecionados o processo de realizar a classificação terá seu atributo de desempenho diretamente afetado.

Ao compararmos ambas abordagens, é possível observar que para grande parte das métricas calculadas os valores são bem semelhantes, tendo somente algumas divergências em casos particulares, isso se dá pela adição do ruído Gaussiano na redução de dimensionalidade diferencialmente privada, assim como também as divergências ocorreram na seleção dos hiperparâmetros, que foram apresentados na Seção 4.

Nas duas abordagens realizadas, temos que, o modelo de Regressão Logística, é o que obtém menor desempenho, mostrando que os dados não são linearmente separáveis, apesar de que ao observarmos a área sobre a curva ROC, nas duas abordagens, temos um resultado bom comparado a outros modelos, o que no caso não o torna melhor, devido a potencialidade das outras métricas.

Para melhor visualização dos resultados já expostos através das métricas, apresentamos também as matrizes de confusão para cada um dos modelos nas duas abordagens explanadas, que podem ser vistas nas Tabelas 3 e 4, contendo os valores de verdadeiros positivos, que são os pacientes com doença de Parkinson testados corretamente, assim como os verdadeiros negativos, no que diz respeito aos pacientes saudáveis. Também temos os valores de falsos positivos e falsos negativos, que consistem respectivamente, em pacientes saudáveis testados com doença de Parkinson e pacientes com a doença ditos saudáveis, sendo este último o pior cenário para o problema em questão.

Um fato interessante a ser citado é que na abordagem não privada, logo após o

Tabela 3. Matriz de confusão - Abordagem não privada

	RL	AGD	AD	SVM	RF	KNN	**PG**
Verdadeiro Positivo	93	136	126	133	138	136	**138**
Verdadeiro Negativo	35	08	17	18	13	14	**16**
Falso Positivo	13	40	31	30	35	34	**32**
Falso Negativo	48	05	15	08	03	05	**03**

Tabela 4. Matriz de confusão - Abordagem privada

	RL	AGD	AD	SVM	**RF**	KNN	PG
Verdadeiro Positivo	93	138	121	137	**137**	135	137
Verdadeiro Negativo	34	04	20	13	**17**	14	12
Falso Positivo	14	44	28	35	**31**	34	26
Falso Negativo	48	03	20	04	**04**	06	04

processo Gaussiano, o modelo *Random Forest* é o que obtém o segundo melhor resultado, analogamente, na abordagem privada, o segundo melhor modelo, após o *Random Forest* é o processo Gaussiano, ou seja, esse resultado em ambos os modelos acabam por divergir devido o ruído adicionado pelo algoritmo Mod_SULQ, o que pode acontecer que em algumas execuções dos algoritmos, nas duas abordagens um mesmo modelo ser considerado ótimo, ou modelos diferentes podem ser os melhores nas abordagens sem privacidade e com privacidade.

6. Conclusão

Os resultados experimentais realizados indicaram que o melhor modelo obtido para a abordagem não privada é o Processo Gaussiano, utilizando a função Racional Quadrática como função de kernel. Já para a abordagem privada, o melhor modelo foi o *Random Forest*, com 137 estimadores, 8 níveis de profundidade máxima e utilizando o índice de Gini como índice de pureza. O trabalho atendeu o objetivo de avaliar diversos modelos para classificação de doença de Parkinson e mostrou que é possível realizar esse processo de modo que os dados dos indivíduos sejam privados, o que dificulta o trabalho de descoberta de informação dos pacientes. Ressalta-se que o que dificulta a classificação de algumas coletas de voz é propriamente a distribuição e comportamento dos dados, o que não permite uma ótima classificação de algumas coletas, além da existência de *outliers*.

Trabalhos futuras envolvem a avaliação de outros algoritmos de aprendizagem, como redes neurais artificias; realizar o processo de treinamento dos dados utilizando o método *Leave-One-Subject-Out*, em que todas as coletas de um paciente é utilizada como conjunto de teste, e o restante das coletas é considerado conjunto de treinamento; testar outros valores dos parâmetros de privacidade, para buscar um limiar relativamente ótimo que garante um nível balanceado entre privacidade e utilidade; além de aplicar outra abordagem de redução de dimensionalidade diferencialmente privado, também proposto em [Chaudhuri et al. 2012].

Referências

Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005). Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART sym-*

posium on Principles of database systems, pages 128–138.

Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Erdogdu Sakar, B., Serbes, G., and Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of parkinson's disease. *PloS one*, 12(8):e0182428.

Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., and Müller, G. (2011). Aspects of privacy for electronic health records. *International journal of medical informatics*, 80(2):e26–e31.

Kaggle (2020). Parkinson's disease (pd) classification. `https://www.kaggle.com/dipayanbiswas/parkinsons-disease-speech-signal-features`. Acessado em 13/07/2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ramani, R. G. and Sivagami, G. (2011). Parkinson disease classification using data mining algorithms. *International journal of computer applications*, 32(9):17–22.

Sakar, C. O. and Kursun, O. (2010). Telediagnosis of parkinson's disease using measurements of dysphonia. *Journal of medical systems*, 34(4):591–599.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., and Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. *Applied Soft Computing*, 74:255–263.