UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE E BIOLÓGICAS

DENISSON AUGUSTO BASTOS LEAL

# PREDIÇÃO DE DISCINESIA EM PACIENTES DA DOENÇA DE PARKINSON USANDO APRENDIZADO DE MÁQUINA

PETROLINA - PE

2021

**DENISSON AUGUSTO BASTOS LEAL**

# PREDIÇÃO DE DISCINESIA EM PACIENTES DA DOENÇA DE PARKINSON USANDO APRENDIZADO DE MÁQUINA

Dissertação apresentada a Universidade Federal do Vale do São Francisco - UNIVASF, Campus Sede, como requisito para obtenção do título de Mestre em Ciências com ênfase na Linha de Pesquisa: Fundamentação Conceitual e Metodologias Inovadoras de Integração em Ciências Ambientais, Tecnologia e Saúde.

Orientador: Profa Dra. Ivani Brys
Coorientador: Prof. Dr. Rodrigo Pereira Ramos

**PETROLINA - PE**

**2021**

Leal, Denisson Augusto Bastos
L435p Predição de discinesia em pacientes da Doença de Parkinson usando aprendizado de máquina / Denisson Augusto Bastos Leal. – Petrolina - PE, 2021.
xi, 57 f. : il. ; 29 cm.

Dissertação (Mestrado em Ciências da Saúde e Biológicas) Universidade Federal do Vale do São Francisco, Campus Petrolina, Petrolina - PE, 2021.

Orientadora: Profª. Drª. Ivani Brys.
Banca examinadora: Profª. Drª. Bruna Del Vechio Koike, Prof. Dr. Ricardo Argenton Ramos.

Inclui Bibliografia.

1. Doença de Parkinson. 2. Aprendizado de Máquina. 3. Discinesia. 4. Parkinson. I. Título. II. Brys, Ivani. III. Universidade Federal do Vale do São Francisco.

CDD 616.833

Ficha catalográfica elaborada pelo Sistema Integrado de Biblioteca SIBI/UNIVASF
Bibliotecária: Adriana Santos Magalhães CRB-4/2275

**UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO**
**PÓS-GRADUAÇÃO CIÊNCIAS DA SAÚDE E BIOLÓGICAS**

## FOLHA DE APROVAÇÃO

DENISSON AUGUSTO BASTOS LEAL

PREDIÇÃO DE DISCINESIA EM PACIENTES DA DOENÇA DE PARKINSON USANDO APRENDIZADO DE MÁQUINA

Dissertação apresentada como requisito para obtenção do título de Mestre em Ciências com ênfase na linha de pesquisa: Fundamentação Conceitual e Metodologias Inovadoras Integradoras em Ambiente, Tecnologia e Saúde, pela Universidade Federal do Vale do São Francisco.

Aprovada em: 13 de setembro de 2021

**Banca Examinadora**

Ivani Brys, Doutora
Universidade Federal do Vale do São Francisco – Univasf

Bruna Del Vechio Koike, Doutora
Universidade Federal do Vale do São Francisco – Univasf

Ricardo Argenton Ramos, Doutor
Universidade Federal do Vale do São Francisco – Univasf

## AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer à minha orientadora Dra. Ivani Brys, meu coorientador Dr. Rodrigo Ramos e a Carla Michele que me ajudaram em toda a jornada deste trabalho.

E a minha família, em especial minha mãe Arnalva Bastos, pelo apoio dado durante todos esses anos.

À Universidade Federal do Vale do São Francisco (UNIVASF), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Programa de Pós-Graduação Ciências da Saúde e Biológicas (PPGCSB), que tornaram essa pesquisa possível.

“Nada é permanente, exceto a mudança.”

**Heráclito**

**RESUMO**

A Doença de Parkinson (DP) é a segunda doença neurodegenerativa mais comum, depois da Doença de Alzheimer, e afeta principalmente pessoas acima dos 60 anos. A DP ainda não tem cura, porém os sintomas podem ser tratados com levodopa, que funciona muito bem nos estágios iniciais da doença. A longo prazo, no entanto, a levodopa está associada ao aparecimento de complicações motoras, denominadas discinesias, em pelo menos metade dos pacientes. A discinesia compromete a funcionalidade do paciente e o manejo farmacológico dos sintomas da DP, e não há, atualmente, medidas efetivas para preveni-las. A Fundação Michael J. Fox desenvolveu um estudo longitudinal com pacientes da DP, o Parkinson Progression Marker Initiative (PPMI), que visa identificar biomarcadores para a progressão da DP, e que reuniu informações motoras, não motoras, neurológicas e sociais de pacientes por cerca de dez anos. Usando técnicas de aprendizado de máquina no banco de dados do PPMI, esse trabalho buscou identificar pacientes da DP que estão em maior risco de desenvolver discinesias. Através da mineração dos dados do PPMI, foram extraídas características clínicas, comportamentais e neurológicas dos pacientes com a DP, que contribuíram para a criação de modelos preditivos. Foram usados os classificadores Multilayer Perceptron, Support Vector Machine, Random Forest, AdaBoost e Regressão Logística, e como medidas de desempenho foram calculadas a acurácia, a curva ROC e a área sob a curva (AUC). Foi realizado também um teste de limitação e alcance do método buscando identificar o menor número de *features* e a maior antecedência possível para previsão. O Random Forest foi o classificador com desempenho mais consistente dentre os testados, e a melhor configuração foi com antecipação média de 9 meses, desvio padrão de 6,1 meses, quando a acurácia chegou a $90,8\%$ e sua AUC ROC a $93,8\%$. Do total de 54 *features* testadas, 16 mostraram-se necessárias e suficientes para essa previsão, sendo as mais importantes, o escore do paciente na parte III da Escala Unificada para Avaliação da DP, a fluência semântica e a dose de medicamentos com ação dopaminérgica. Foi observado também que o desempenho do classificador caiu, significativamente, à medida que o tempo de antecedência em relação ao início da discinesia foi aumentado. Isso pode ter acontecido devido ao fato de que com o aumento de tempo os sintomas ficam menos aparentes, dificultando a diferenciação dos pacientes que não desenvolvem discinesia, mas também a limitação temporal imposta levou a um aumento na quantidade de *missing values* e uma queda na quantidade de registros de pacientes. Este foi um estudo interdisciplinar e pioneiro, no qual demonstramos que é possível identificar antecipadamente pacientes da DP que estão em maior risco de desenvolver discinesias através de aprendizado de máquina. Nossos resultados contribuem para o desenvolvimento de medidas preventivas que visam evitar ou retardar o início das discinesias, e podem ainda servir para orientar a terapia dopaminérgica na prática clínica.

**Palavras-chave**: Aprendizado de Máquina. Parkinson. Discinesia, PPMI.

**ABSTRACT**

Parkinson's Disease (PD) is the second most common neurodegenerative disease after Alzheimer's Disease, and it mainly affects people over 60 years. PD is not curable yet, but the symptoms can be treated with levodopa, which works very well in the early stages of PD. In the long term, however, levodopa has been associated with motor complications, known as dyskinesias, in more than half of patients. Dyskinesia compromises the patient's motricity and response to PD pharmacological treatment, and currently there are no effective measures to prevent it. The Michael J. Fox Foundation performed a longitudinal study of PD patients, the Parkinson Progression Marker Initiative (PPMI), which aimed to identify biomarkers for PD progression, and has gathered motor, non-motor, neurological, and social patients information for about ten years. Using machine learning techniques in the PPMI database, this work sought to identify in advance PD patients who were at increased risk of developing dyskinesias. Patient's clinical, behavioral, and neurological characteristics were mined from PPMI's database in order to create predictive models. Multilayer Perceptron, Support Vector Machine, Random Forest, AdaBoost and Logistic Regression classifiers were used. Accuracy, ROC curve and area under the curve (AUC) were calculated as performance measures. A test of limitation and scope of the method was also carried out, seeking to identify the smallest number of features and the longest possible advance for forecasting. Random Forest had the most consistent performance amongst the tested methods. The best configuration had a mean anticipation of 9 months and standard deviation of 6.1 months, reaching 90.8% of accuracy and AUC ROC of 93.8%. Out of 54 total tested features, 16 were necessary and sufficient for this prediction. The most important were: the patient's score in part III of the Unified Parkinson's Disease Rating Scale, the semantic fluency and the dose of dopaminergic medications. It was also observed that the classifier's performance dropped significantly as time in relation to dyskinesia onset increased. Such a result may due to the fact that the symptoms were less apparent at that time point, making it difficult to correctly differentiate the patients. We can not ruled completely out that increased the amount of missing values and reduced the number of patient records. This was an interdisciplinary and pioneer study that used machine learning to demonstrate that it is possible to identify in advance PD patients with increased risk of developing dyskinesias. Our results contribute to the development of measures aiming at preventing or delaying the onset of dyskinesias, and may also help to guide PD dopaminergic therapy in the clinical practice.

**Keywords**: Machine Learning. Parkinson. Dyskinesia. PPMI.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AUC | Area Under the ROC Curve |
| COMT | Catecol-O-metil Transferase |
| DP | Doença de Parkinson |
| FDA | Food and Drug Administration |
| IA | Inteligência Artificial |
| ICOMT | Inibidores da catecol-O-metiltransferase |
| JLOT | Teste de Orientação da Linha |
| LEDD | Levodopa Equivalent Daily Dosage |
| LLE | Locally Linear Embedding |
| MAO-B | Monoamine Oxidase B |
| ML | Machine Learning |
| MLP | Multilayer Percepteron |
| NMDA | N-metil-D-aspartato |
| PCA | Principal Component Analysis |
| PDBP | Parkinson's Disease Biomarker Program |
| PPMI | Parkinson Progression Markers Initiative |
| RNA | Redes Neurais Artificiais |
| ROC | Receiver Operating Characteristics |
| SVM | Support Vector Machine |

**SUMÁRIO**

# 1 INTRODUÇÃO

A Doença de Parkinson (DP) é a segunda doença neurodegenerativa mais comum, depois da Doença de Alzheimer, afetando 1% da população mundial acima dos 60 anos de idade, e 3% acima dos 80 anos (BALESTRINO; SCHAPIRA, 2020). De acordo com a projeção realizada por Dorsey et al. (2007), estima-se que, nos cinco países mais populosos da Europa somados com os dez mais populosos do mundo, o número de pacientes da DP com mais de 50 anos alcance os 8.67 milhões em 2030. No Brasil, o aumento estimado é de 112%, com o número de pessoas diagnosticadas com a DP passando de 160 mil para aproximadamente 340 mil no mesmo período (DORSEY et al., 2007).

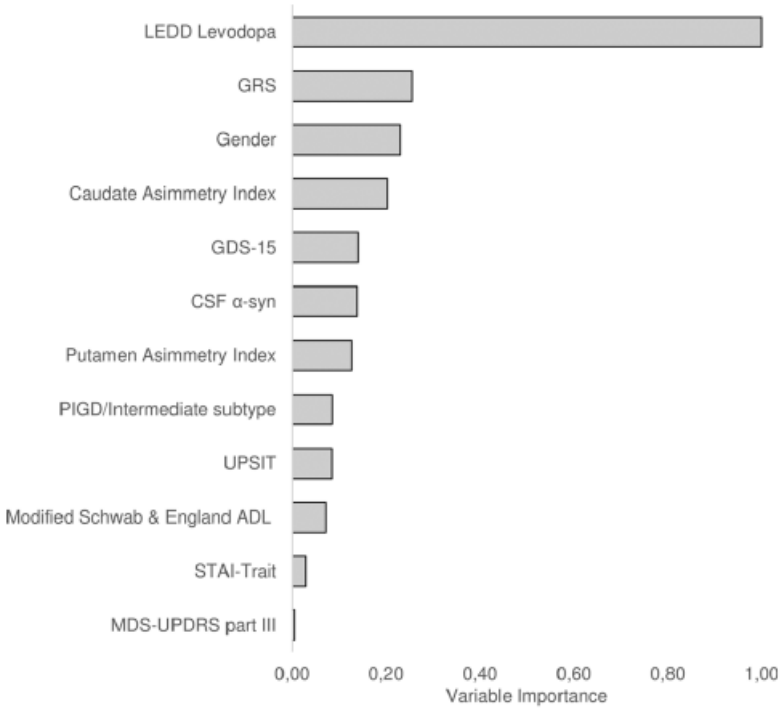
Atualmente não há cura para a DP, e o tratamento padrão ouro tem como objetivo aliviar os sintomas da doença através do fármaco levodopa, que é um precursor dopaminérgico (BALESTRINO; SCHAPIRA, 2020). A levodopa é efetiva em aliviar os sintomas em curto prazo, mas em longo prazo está associada ao aparecimento de complicações motoras conhecidas como discinesias, que comprometem a funcionalidade do paciente e o manejo farmacológico dos sintomas parkinsonianos. Após cinco anos de uso da levodopa, cerca de 50% dos pacientes desenvolvem discinesia, e após dez anos, 80% (ZESIEWICZ; SULLIVAN; HAUSER, 2007). Atualmente, não é possível prevenir o aparecimento de discinesia, pois quando um paciente é diagnosticado com a DP e inicia o tratamento com levodopa, não é possível saber se o mesmo desenvolverá discinesia e nem quando isso acontecerá.

O Parkinson Progression Marker Initiative (PPMI) é um estudo longitudinal realizado pela Fundação Michael J. Fox, que tem como objetivo identificar biomarcadores da doença para melhorar a sua compreensão e fornecer ferramentas que colaborem com o sucesso de testes terapêuticos (MAREK et al., 2011). Este estudo acompanhou centenas de pacientes da DP por um período mínimo de três anos, durante os quais avaliações neurológicas, clínicas, exames de imagem e coletas de amostras biológicas foram sistematicamente realizadas. O banco de dados PPMI representa, portanto, um conjunto único de dados referentes ao acompanhamento longitudinal de pacientes da DP, através do qual, é possível identificar marcadores que podem informar sobre a progressão da doença e o aparecimento futuro de discinesias nos pacientes em tratamento com levodopa.

Recentemente, Eusebi et al. (2018) analisaram 39 características de 423 pacientes com DP do PPMI, através de uma análise de regressão multivariada de Cox, e concluíram que a exposição cumulativa à levodopa, o gênero feminino, a gravidade do comprometimento motor e funcional, o fenótipo clínico não dominante de tremor, o risco genético, a ansiedade e a assimetria do caudado são fatores de risco associados ao surgimento de discinesias. Neste mesmo estudo, outra análise feita com o Random Survival Forest, que pode ser

vista na Figura 1, encontrou variáveis que o classificador considerou mais importantes para identificar pacientes com discinesia. Além das já citadas, destacam-se os índices de depressão avaliados através da escala de depressão GDS-15, a $\alpha$-sinucleína e o índice de assimetria do putamen (EUSEBI et al., 2018).

**Figura 1** – Cálculo de importância dos fatores de risco na discinesia.




**Fonte:** Eusebi et al. (2018)

Os recentes esforços para a identificação dos fatores de risco para o aparecimento de discinesias abriram oportunidades para o desenvolvimento de ferramentas potencialmente capazes de identificar, com anos de antecedência, pacientes da DP que estão em maior risco para desenvolver discinesias no futuro. As técnicas de aprendizado de máquina (ML, do inglês *Machine Learning*) são formas automática através da qual algoritmos computacionais aprendem novas informações por meio de experiências passadas.

Solana-Lavalle e Rosas-Romero (2021) conduziram um experimento usando diversos tipos de modelos de aprendizado de máquina para diagnosticar pacientes com a doença de Parkinson a partir de uma base de imagens de ressonância magnética, também fornecidos pelo banco de dados do PPMI. Esse trabalho passou por diversas etapas intermediárias como identificação de áreas de interesse do cérebro, extração e seleção de características e separação por gênero, através das quais os autores obtiveram excelentes resultados com o classificador Naive Bayes para os homens e a Regressão Logística para as mulheres, chegando a uma acurácia de $99,01\%$ e $96,97\%$, respectivamente (SOLANA-LAVALLE;

ROSAS-ROMERO, 2021).

Ao invés de identificar padrões manualmente e depois automatizar o processo escrevendo um programa de computador, a aplicação de técnicas de aprendizado de máquina a um conjunto de dados com um certo padrão possibilita criar relações entre os dados e obter uma resposta desejada com um conhecido grau de exatidão mesmo que não seja inicialmente perceptível (DAS; BEHERA, 2017). Este, portanto, é um estudo interdisciplinar e pioneiro que visa encontrar um classificador, através da aplicação de técnicas mineração de dados no banco de dados PPMI e de ML, que seja capaz de identificar de forma otimizada e com antecedência o risco de uma pessoa desenvolver discinesias no futuro.

## 1.1 OBJETIVO GERAL

Identificar, através de técnicas de aprendizado de máquina, características comportamentais e neurológicas associadas ao desenvolvimento de discinesias em pacientes da Doença de Parkinson.

## 1.2 OBJETIVOS ESPECÍFICOS

- Selecionar no banco de dados PPMI os pacientes da DP e identificar os que desenvolveram discinesias ao longo do estudo.
- Caracterizar os pacientes selecionados no que diz respeito aos sintomas neurológicos, motores, não motores, características sociodemográficas e da doença.
- Identificar características dos pacientes que mais contribuem para a sua classificação nos grupos com e sem discinesia.
- Construir modelos de aprendizado de máquina que identifiquem os pacientes que estão em maior risco de desenvolver discinesia no futuro.
- Avaliar limitações e o alcance do método em relação à quantidade mínima de características e à antecedência para classificação da discinesia.

# 2 FUNDAMENTAÇÃO TEÓRICA

O uso de aprendizado de máquina tem sido cada vez mais ampliado nas pesquisas da ciência da computação para outras áreas como a economia, saúde e engenharia, dentro e fora da academia. Hoje, no meio científico, é utilizado principalmente para processar grandes volumes de dados, encontrar relações entre variáveis e colaborar para o entendimento do problema abordado (ROSCHER et al., 2020). A saúde, em especial, tem vários pontos onde a tomada de decisão pode ser auxiliada pelo uso de Inteligência Artificial (IA), tornando o processo mais exato e com um maior número de compartilhamento de experiências. Cada dia mais pesquisas envolvendo IA na Neurociência vêm sendo publicadas, com trabalhos com mapeamento e simulações de áreas do cérebro e no entendimento e diagnóstico de doenças, como a Doença de Parkinson feita neste estudo. Nesse capítulo, falaremos sobre conceitos fundamentais para o entendimento do problema de pesquisa da presente dissertação.

## 2.1 APRENDIZADO DE MÁQUINA

Aprendizado de máquina é uma subárea da Inteligência Artificial na qual o conhecimento não é explicitamente programado. Então, ao invés de fornecer um conjunto de regras, uma forma mais genérica é utilizada, na qual são fornecidos dados de exemplos da situação e o problema é explorado, corrigido e aprendido pelo sistema (JANIESCH; ZSCHECH; HEINRICH, 2021). Existem várias distinções entre ML e a modelagem estatística, segundo Dangeti (2017), de difícil diferenciação, sendo algumas delas: a formalização matemática do relacionamento entre as variáveis que é feito na modelagem estatística e em ML não; na estatística é necessário assumir a forma da curva antes de realizar o ajuste do modelo enquanto no ML os padrões são aprendidos automaticamente, por mais que sejam complexos; em ML, é necessária uma divisão extra dos dados chamada de validação que é utilizada antes dos testes; entre outras características.

Os métodos de ML podem ser divididos em três categorias: supervisionado, não supervisionado e por reforço. O aprendizado supervisionado pode ser visto como o aprendizado de uma função, em que o dados de treinamento possuem as variáveis de entrada e a saída, que também chamada por alguns autores de valor resposta, rótulo ou variável alvo. No não supervisionado, os dados não possuem o rótulo e o aprendizado é feito por associação ou agrupamento. No aprendizado por reforço, o algoritmo tenta fazer algumas ações e é punido por um comportamento errado ou recompensado por trabalhos corretos (GÉRON, 2019). Nesse trabalho, daremos ênfase ao aprendizado supervisionado pela característica dos dados.

### 2.1.1 Classificadores

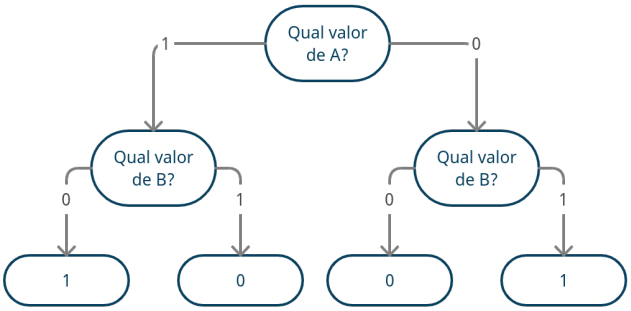
Exitem diversos tipos de classificadores, alguns têm um embasamento estatístico mais sólido, outros têm inspiração na inteligência humana, no comportamento da natureza, entre outros. Nesse projeto, usamos sete classificadores: Regressão Logística, Árvore de Decisão, Multilayer Perceptron, Support Vector Machine, Random Forest, Gradient Boosting e AdaBoost.

Partindo de classificadores mais próximos da estatística, a Regressão Logística possui um princípio semelhante à Regressão Linear Múltipla diferindo na forma em que o resultado é expresso, que é uma probabilidade de o resultado pertencer a uma classe alvo (BRUCE; BRUCE, 2017). A Regressão Logística pode ser expressa pela Equação (1), na qual os parâmetros $\beta$ são estimados pela máxima verossimilhança, $X$ são as variáveis independentes, a função logarítmica ($log$) é usada para que o resultado possa variar de $-\infty$ (infinito negativo) a $+\infty$ (infinito positivo) e $odds$ representa as chances da variável dependente ocorrer (DANGETI, 2017).

$$log(odds) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_n * X_n \quad (1)$$

Um dos modelos de mais fácil entendimento do funcionamento e fácil explicabilidade dos resultados é a Árvore de Decisão. Após o treinamento, o seu resultado pode ser explicado de uma forma simples porque ela é composta por pontos de perguntas e a resposta de cada pergunta serve como um guia até chegar no resultado do modelo (DANGETI, 2017). A Figura 2 mostra um exemplo de Árvore de Decisão de uma porta XOR - que possui duas entradas A e B, quando A é igual a B o resultado é 0 e quando são diferentes o resultado é 1 - que inicia perguntando sobre a entrada A; caso seja 0 pergunta pela entrada B onde a resposta é 1 caso seja 1 e 0 caso seja 0. Se A for 1, a entrada B também é verificada, mas dessa vez com resposta contrária.

**Figura 2** – Árvore de decisão para uma porta XOR com entradas em A e B.

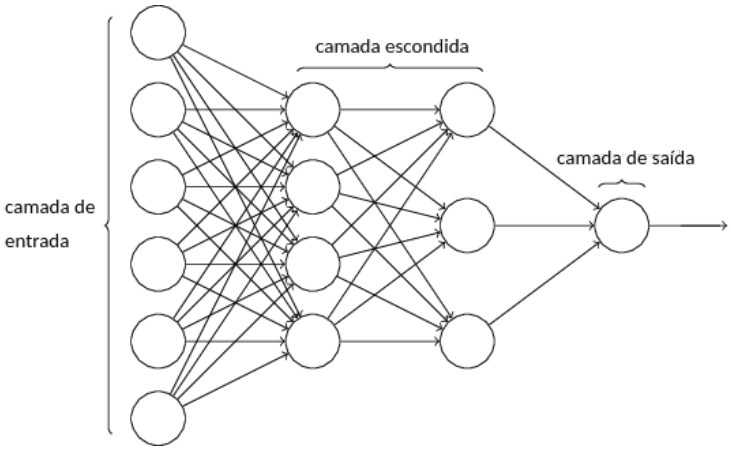



**Fonte:** O autor

Apesar da Figura 2 mostrar um exemplo de um modelo de classificação com dados categóricos na entrada, a Árvore de Decisão também pode dar respostas de regressão e receber dados contínuos como entrada. Além disso, ela trabalha de forma não paramétrica (podendo ter como exceção a altura máxima) e não precisa ter a mesma altura em todos os pontos (DANGETI, 2017).

Multilayer Perceptron (MLP) é um tipo de Rede Neural Artificial (RNA) composta por Perceptrons que, por sua vez, são os antecessores mais populares dos RNAs. O Percepton pode ser representado matematicamente por $y = W^T X + b$, onde $X$ é a entrada, $W$ é a matriz de pesos de cada entrada, $b$ é o *bias* e $y$ é a saída da rede. No caso de problemas de classificação existe um limiar, que quando $y$ é maior que ele, o resultado é 1 e quando menor ou igual, resulta em 0 (UNPINGCO, 2016). O MLP é formado por um conjunto de Perceptrons que formam uma topologia com três partes: a camada de entrada, as camadas escondidas e a camada de saída. Cada camada possui um ou mais Perceptrons, como pode ser visto na Figura 3. O MLP usa a técnica *back propagation* para seu aprendizado supervisionado (KINGE; GAIKWAD, 2018).

**Figura 3** – Ilustração do Multilayer Perceptron e suas camadas internas.

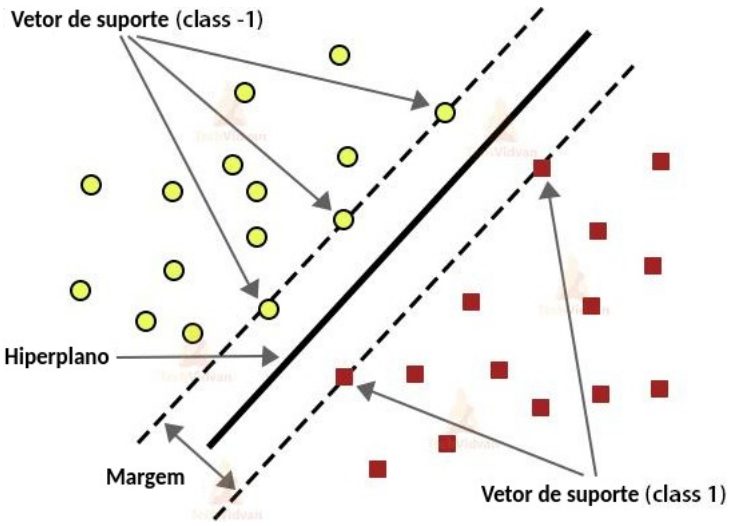



**Fonte:** Adaptado de Liashchynskyi (2019)

O Support Vector Machine (SVM) é mais um tipo de modelo de ML que usa o aprendizado supervisionado e cria um hiperplano para separar duas classes de forma otimizada. A Figura 4 ilustra um hiperplano de separação das classes, definida pela equação $\omega\varphi(x) + b = 0$, onde $\omega$ representa o vetor normal ao hiperplano, e os dois vetores de suporte equidistantes. Cada ponto no plano representa um elemento de uma das classes a serem consideradas. O classificador escolhe a classe usando as características como coordenadas para saber de que lado do hiperplano o ponto se encontra. É importante ressaltar que apesar da ilustração estar em duas dimensões, o hiperplano usado pode ter $N$ dimensões, que representam as características e vão depender da complexidade do problema (DAS;

BEHERA, 2017).

**Figura 4** – Support Vector Machine.




**Fonte:** Adaptado de Agrawal (2020)

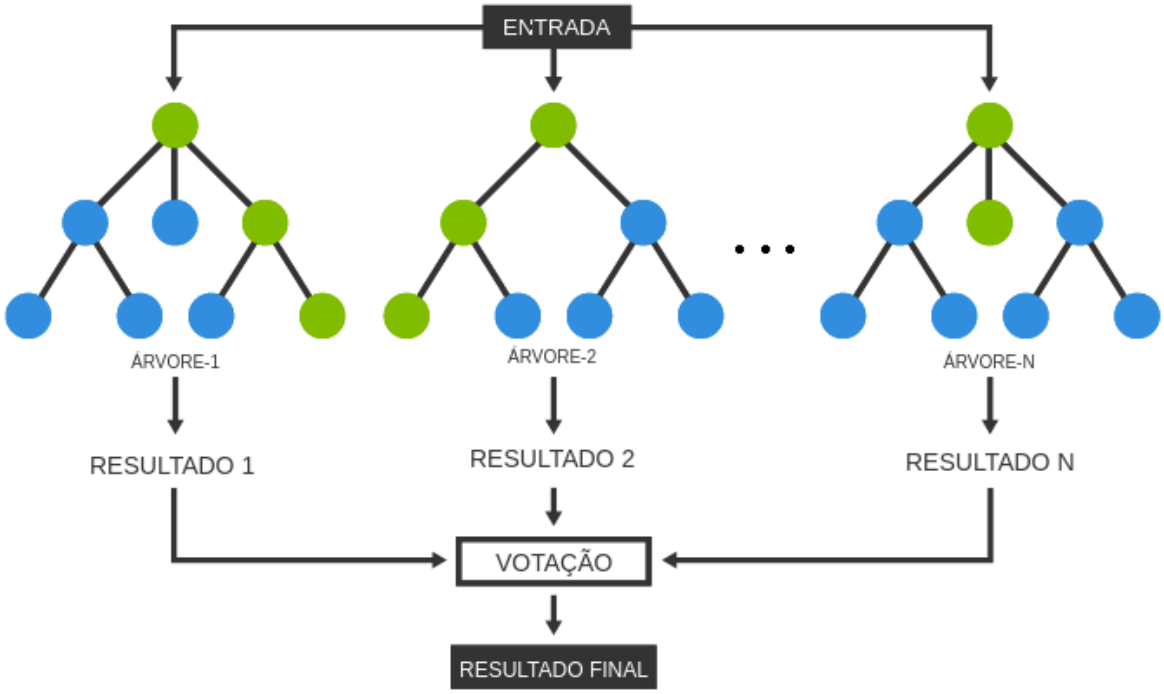
Um ponto interessante do SVM é que ele dá ênfase para os pontos mais próximos da região de fronteira, que são os vetores de suporte. Em outras palavras, os demais pontos são irrelevantes para a construção do modelo. Isso é muito interessante em alguns casos, porque pode reduzir a variância do modelo e demandar um conjunto de dados de treinamento menor, já que geralmente uma pequena parte está próximo dessa região (UNPINGCO, 2016).

Métodos com *ensemble* são muito usados desde os anos 90 e são basicamente conjuntos de vários modelos diferentes ou criados de forma diferente, em que os resultados podem ser vistos como um consenso entre os votos de cada um (CHEN et al., 2020). Uma analogia que pode ser feita é imaginar um conjunto de vários especialistas, no qual cada um chega ao seu diagnóstico do problema e por meio de uma votação o grupo chega a uma conclusão.

Random Forest é um exemplo de *ensemble* com diversas Árvores de Decisão e, para ter árvores crescendo de forma diferente, dois processos são realizados: o treinamento feito com uma amostra de dados diferente e com subconjuntos das características em diferentes árvores (SAGI; ROKACH, 2018). A Figura 5 ilustra o funcionamento geral do Random Forest, onde N árvores são treinadas e usadas para realizar a classificação, em seguida, usando todos os resultados, é aplicado uma votação, onde a classe com o maior número de votos será a saída final do classificador.

AdaBoost é outro exemplo que, como descrito por Kinge e Gaikwad (2018), treina vários modelos simples e com desempenhos pouco superiores a uma decisão aleatória,

**Figura 5** – Ilustração do Random Forest.




**Fonte:** Adaptado de Tibco (2021)

combinando os resultados usando maioria ou a soma para ter o resultado final. O Gradient Boosting, por sua vez, é um exemplo de ensemble que usa Árvores de Decisão e tem como principais diferenças a otimização de uma função de perda e o uso das falhas dos demais classificadores para treinamento da Árvore seguinte com objetivo de diminuir a função de perda (KLUG et al., 2020).

Os classificadores possuem algumas limitações, normalmente sendo tão robustos quanto os dados de treinamento. Isso refere-se tanto à precisão da medição quanto ao que o dado medido representa no contexto do problema. Medições feitas em uma região geográfica ou com um grupo específico também pode influenciar, pois a amostra coletada pode não ser suficientemente representativa e apresentar problemas na generalização. Alguns modelos, principalmente de aprendizado profundo, podem precisar de uma enorme quantidade de dados para seu treinamento e isso pode ser visto como um limitante para sua adoção. Em problemas da saúde, é importante saber justificar a decisão tomada e modelos mais complexos são como uma caixa preta, onde é quase impossível extrair o motivo da decisão (PAI; PAI, 2021).

### 2.1.2 Redução de dimensionalidade

Em alguns problemas de aprendizado de máquina, o número de características pode ser muito grande deixando o treinamento difícil de convergir e com isso exigir mais

tempo e poder de processamento. Existem algumas formas de contornar esse problema, que também é conhecido como a maldição da dimensionalidade. A forma mais popular de lidar com esse problema é usar a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*), porém existem outras que não serão abordadas, como Kernel PCA e Locally Linear Embedding (LLE) (ABAS et al., 2020).

A PCA é um método de redução de dimensionalidade que trabalha de forma não supervisionada (AROWOLO; ADEBIYI; ADEBIYI, 2020). Ela usa uma técnica de decomposição em valores singulares que é capaz de decompor uma matriz em um produto escalar de três matrizes $U \cdot \Sigma \cdot V^T$ e a matriz $V$ terá os componentes principais. Uma vez encontrada essa matriz, para reduzir a dimensionalidade, basta projetar a matriz original em um hiperplano formado pelas $n$ primeiras dimensões na matriz $V$, sendo $n$ o número de dimensões desejadas (GÉRON, 2019).

### 2.1.3 Medidas de desempenho

Depois da criação dos modelos, é importante conhecer sua eficiência e se eles têm capacidade de resolver o problema em questão com desempenho superior aos modelos já empregados ou, caso não haja referências, ao desempenho puramente aleatório. Uma das formas mais comuns de medir o desempenho de problemas de classificação é a acurácia, que consiste na taxa de classificações corretas pelo número total de testes (BALLABIO; GRISONI; TODESCHINI, 2018).

A Curva Característica de Operação do Receptor (ROC, do inglês *Receiver Operating Characteristics*), por sua vez, é uma análise gráfica da performance, usada normalmente em classificações binárias (FAWCETT, 2006). A curva ROC é construída utilizando a taxa de acertos positivos pelo total de positivos no eixo $y$ e a taxa de erros em classificações positivas pelo total de negativos no eixo $x$ (FAWCETT, 2006). A área sob a curva ROC (AUC, do inglês *Area Under the ROC Curve*) é uma forma eficiente de comparar classificadores utilizando a curva ROC e consiste no cálculo da área delimitada pela curva que varia entre 0 e 1, sendo o melhor resultado aquele mais próximo de 1 (FAWCETT, 2006).

### 2.1.4 Ajuste de Modelos

Um classificador pode ter mais de uma dezena de hiperparâmetros e a busca pela combinação com melhor desempenho pode ser árdua e manualmente exaustiva. O GridSearch é uma forma bastante utilizada para resolver esse tipo de problema (GÉRON, 2019). Ele tem o objetivo de encontrar um conjunto ótimo de hiperparâmetros de um classificador tomando como base os dados de treinamento. Ele encontra o melhor resultado do modelo fazendo uma varredura completa, passando por todas as combinações de parâmetros disponíveis (ABAS et al., 2020).

Dependendo da quantidade e da variação dos hiperparâmetros, o GridSearch pode levar muito tempo para conclusão. Como alternativa, existe o RandomizedSearch que possui o mesmo princípio, mas com um espaço de busca limitado. De acordo com o limite recebido, o RandomizedSearch busca aleatoriamente combinações de hiperparâmetros que não se repetem, e realiza a busca da melhor combinação na amostra selecionada (GÉRON, 2019).

Ao falar de ajuste de modelo, a primeira impressão é que o modelo estará bem ajustado para esses dados usados como entrada, mas isso não garante que continuará ajustado para os dados que ele não conhece. Pensando nisso, o *dataset* utilizado é usualmente dividido em três conjuntos: treinamento, validação e teste. O conjunto de treinamento é usado para ajustar os parâmetros do modelo ao problema abordado. O conjunto de validação pode ser usado para testar diversas combinações de hiperparâmetros e até fazer mais de um treinamento para garantir que o modelo não caiu em um mínimo local. Para algumas aplicações o conjunto de validação pode não ser necessário e esse conjunto passa a integrar o conjunto de treinamento. O conjunto de testes é fundamental para solucionar esse problema, pois ele não é usado em nenhuma situação anterior e só é feito a classificação uma vez, no final, quando o modelo já está pronto. Com o conjunto de teste, usando as medidas de desempenho, avaliamos a generalização do classificador para dados não vistos pelo modelo, tornando mais próximos dos dados reais (GÉRON, 2019).

## 2.2 DOENÇA DE PARKINSON

A Doença de Parkinson (DP) é uma condição neurodegenerativa sem cura, caracterizada pela perda de neurônios dopaminérgicos na via nigroestriatal. O diagnóstico é feito de forma clínica normalmente depois dos sintomas motores aparecerem. Quando isso acontece, cerca de 60% dos neurônios dopaminérgicos já foram perdidos, o que dificulta muito o tratamento. Entre os seus sintomas estão tremor em repouso, rigidez, bradicinesia, acinesia, instabilidade postural e outros sintomas motores e não motores (BALESTRINO; SCHAPIRA, 2020).

A DP atinge 0,3% da população e, mesmo com etiologia da doença para maioria dos casos sendo desconhecida, a idade é considerada um fator de risco, pois para pessoas com idade acima de 60 anos a incidência é de 1% e para pessoas acima de 80 anos de idade, a incidência sobe para 3%. Outro fator associado ao surgimento da doença é a vida rural e o contato com pesticidas e com outras substâncias. Algumas substâncias têm sido inversamente relacionadas com a incidência da doença, como o cigarro, café, bloqueadores dos canais de cálcio e estatinas (BALESTRINO; SCHAPIRA, 2020).

### 2.2.1 Tratamento farmacológico da DP

Desde a descoberta do papel da dopamina para o controle motor, na década de 1960, o tratamento farmacológico da DP tem sido feito principalmente através da reposição deste neurotransmissor. O medicamento padrão ouro para o tratamento é a Levodopa, que controla os sintomas motores, mas não a progressão da doença. O uso da Levodopa pode causar alguns efeitos colaterais como náuseas, hipotensão, sonolência, confusão, alucinações, hipersexualidade, compras compulsivas e aumento da propensão aos jogos de azar (BALESTRINO; SCHAPIRA, 2020).

Além disso, o uso prolongado da Levodopa está associado a complicações motoras severas conhecidas como discinesia. Discinesia induzida por levodopa consiste em movimentos involuntários de maior amplitude e mais difíceis de controlar que os próprios sintomas da DP, aumentando o desconforto do paciente assim como o risco de queda. Acredita-se que as complicações estejam ligadas à estimulação dos receptores de dopamina em intervalos de tempo discretos pela reposição farmacológica de dopamina, diferente do fornecimento contínuo natural. Essas complicações motoras podem ter relação com a dose diária de levodopa, gravidade da neurodegeneração dopaminérgica, sexo feminino, baixo peso, entre outros fatores (BALESTRINO; SCHAPIRA, 2020).

Há algum tempo, tentativas de controlar ou retardar o uso de levodopa vêm sendo empreendidas, porém o aumento da dose é necessário para controlar os sintomas motores à medida que a doença progride. Importante notar que o processo degenerativo subjacente à DP é progressivo e, com o avanço do mesmo, menos neurônios dopaminérgicos estão disponíveis para fazer a conversão da levodopa em dopamina, o que resulta na inevitável necessidade de doses maiores para alcançar algum benefício terapêutico. Estima-se que 75% dos pacientes com a DP acabarão por experimentar algum tipo de discinesia após dez anos de uso de levodopa. Hoje, a opção mais eficaz de reduzir a incidência de discinesia é uma intervenção cirúrgica para implante de eletrodos de estimulação cerebral profunda, pois com isso o paciente acaba por necessitar de doses menores de levodopa. Importante notar, porém, que somente 10% dos pacientes geralmente atendem aos critérios para essa cirurgia (CENCI et al., 2020).

Outros medicamentos também são utilizados para o alívio dos sintomas da DP, entre eles os Agonistas Dopaminérgicos. Embora esses medicamentos estimulem os receptores dopaminérgicos diretamente, são menos eficazes que a levodopa e também são associados ao risco de desenvolvimento de discinesias. Além disso, têm outros efeitos colaterais como edema nas pernas, sonolência diurna excessiva e mais distúrbios do controle dos impulsos. Geralmente os agonistas dopaminérgicos são utilizados na fase inicial da doença ou em conjunto com a Levodopa (BALESTRINO; SCHAPIRA, 2020).

Os inibidores da Monoamine Oxidase B (MAO-B) são outros medicamentos também

utilizados na fase inicial da doença ou em conjunto com a levodopa no tratamento dos sintomas da DP. Estes medicamentos atuam diminuindo o metabolismo da dopamina e com isso prolongam a estimulação dos receptores dopaminérgicos. Os Inibidores da Catecol-O-metil Transferase (COMT), por sua vez, aumentam a meia-vida da Levodopa, e são usados com a finalidade de diminuir as flutuações motoras comuns ao tratamento com Levodopa (BALESTRINO; SCHAPIRA, 2020).

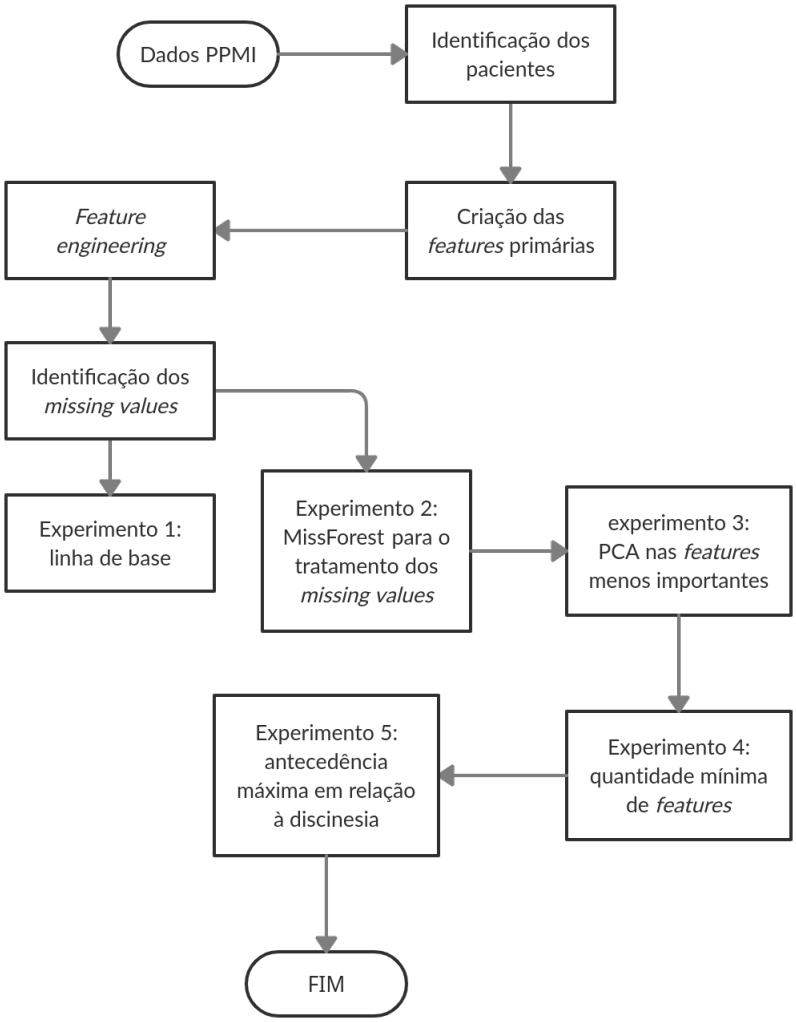
Até então, temos falado de medicamentos que ajudam a diminuir os sintomas motores da DP, porém não podemos deixar de citar a Amantadina que é o único fármaco atualmente aprovado pela Food and Drug Administration (FDA - EUA) para o controle de discinesias induzidas por Levodopa. A amantadina é um antagonista de receptores glutamatérgicos N-metil-D-aspartato (NMDA), e embora alguns estudos tenham demonstrado seu potencial em reduzir sintomas discinéticos, seu uso está associado a diversos efeitos colaterais, como alucinações e outras alterações perceptivas (KUHN; MÜLLER, 2020). Evidências experimentais recentes têm demonstrado que agonistas setoronérgicos, específicos para os receptores 5HT1a podem também aliviar sintomas de discinesia (BRYS et al., 2018). No entanto, o aparecimento de movimentos involuntários anormais continua sendo o grande desafio para o manejo farmacológico dos sintomas da DP em longo prazo.

Diante do que foi exposto, percebe-se que a discinesia é uma característica limitadora para pacientes com DP e sua detecção precoce pode ser de extrema importância para tornar a vida do paciente menos traumática. A aplicação de ML no processo de detecção do risco de desenvolvimento de discinesia torna-se, desta forma, uma problema altamente relevante.

# 3 MATERIAIS E MÉTODOS

Esse capítulo apresentará os passos e procedimentos realizados para alcançar os objetivos propostos e a Figura 6 guiará a passagem por todas as etapas destacando suas conexões.

**Figura 6** – Fluxograma das etapas do estudo.




**Fonte:** O autor

Como ilustrado no fluxograma da Figura 6, inicialmente, foram obtidos os dados dos pacientes, dos quais foram extraídas as características (*features*) que foram usadas como entrada para os classificadores. Foram criadas duas ramificações no fluxo para o processo de classificação. Uma versão básica utilizando as formas mais comuns de pré-processamento, denominado Experimento 1 teve o objetivo de ser utilizado como linha de

base para comparação com as melhorias futuras. A segunda ramificação engloba os diversos experimentos empregados no trabalho, descritos como segue. O Experimento 2 realizou uma melhoria usando o MissForest para preenchimento dos valores faltantes (*missing values*). O Experimento 3, além de usar as melhorias do MissForest, incrementou com o uso da PCA nas características menos importantes.

Os Experimentos 4 e 5, por sua vez, seguiram uma lógica diferente dos demais, pois serviram para testar as limitações e o alcance desse método. Então, o Experimento 4 teve o objetivo de descobrir a quantidade mínima de recurso para se obter um desempenho aceitável e, por fim, o Experimento 5 buscou descobrir o tempo máximo para o qual seria possível obter uma predição relativamente boa. Todas as etapas indicadas no fluxograma serão melhor detalhadas nas seções a seguir.

## 3.1 DESCRIÇÃO DOS DADOS

Neste estudo, analisamos o banco de dados PPMI disponibilizado pela Fundação Michael J. Fox, que contém informações longitudinais de pacientes diagnosticados com a DP com ou sem uma das seguintes mutações genéticas: LRRK2, GBA, ou SNCA. A coleta destes dados foi realizada durante visitas regulares dos participantes aos centros de pesquisa participantes do estudo. Os dados consistem de informações genéticas, sócio-demográficas, comportamentais, neurológicas, entre outros, coletadas através de exames de laboratório, exames de imagem, aplicação de escalas e questionários que podem ser preenchidos pelo paciente, pelo cuidador ou familiar. Este banco é disponibilizado para pesquisadores da área mediante solicitação direta à Fundação.

De acordo com as informações acessadas, o PPMI iniciou a coleta dos dados em 2010, e a última coleta acessada para o presente estudo ocorreu em 2020. O banco conta com dados de 2.254 voluntários acompanhados em diferentes períodos e etapas da doença e de voluntários saudáveis, distribuídos em 141 planilhas.

### 3.1.1 Participantes

Na prática de coleta de dados, principalmente de estudos longitudinais, é comum iniciar a coleta com pacientes e perder o contato com os mesmos, além de, em muitas situações, os participantes iniciarem sua participação no estudo já com a doença em estágio avançado. Diante disso, foi necessário aplicar alguns critérios mínimos para incluir os pacientes nesse estudo, conforme listado a seguir:

1. Possuir diagnóstico da Doença de Parkinson de acordo com os registros na planilha “PD_features”;

2. Estar nos grupos 1 ou 5 da tabela “Screening_Demographics”, que correspondem a pacientes com a DP. No grupo 5, além da DP, foi identificada a presença de alguma das seguintes mutações genéticas: LRRK2, GBA ou SNCA;

3. Ter realizado pelo menos uma avaliação antes do aparecimento da discinesia, de acordo com o registro na coluna “NUPDRS4” da tabela “MDS_UPDRS_Part_IV” que mostra a quantidade de tempo gasto com discinesias. Pacientes que apresentavam discinesia no início do estudo foram excluídos.

4. Não ter registro de uso do medicamento Amantadine, na tabela “Concomitant _Medications”, no início das avaliações, para pacientes que não pontuaram para discinesia. Este medicamento é usado para aliviar discinesias e pode mascarar a incidência das mesmas.

5. Possuir mais de dois terços dos dados das avaliações utilizadas presentes no intervalo de tempo selecionado.

Como resultado desses filtros teremos somente pacientes diagnosticados com a DP que iniciaram o acompanhamento sem discinesia. Alguns desses pacientes ao longo do estudo desenvolveram discinesia e outros não, com isso formam-se dois grupos sendo essa nossa variável alvo.

### 3.1.2 Intervalo entre a coleta e o aparecimento da discinesia

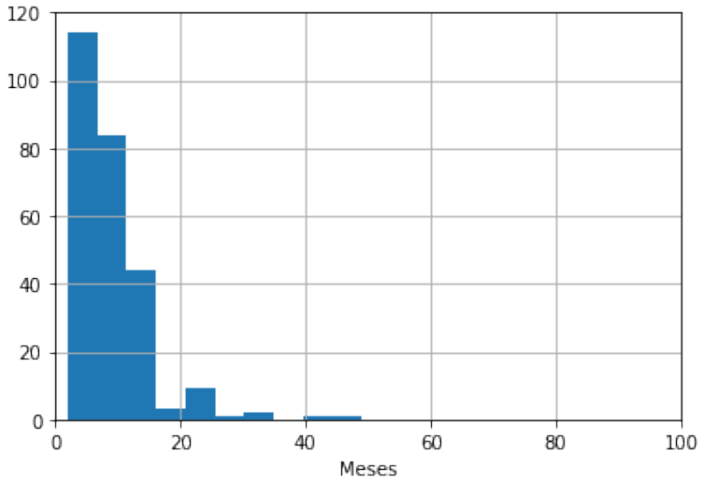
A coleta de informação do PPMI é realizada durante visitas, feitas em média anualmente, e são denominadas de BL (linha de base), V01 (visita 1), V02 (visita 2) e assim sucessivamente. Mas nem todas as informações são coletadas em todas as visitas. Alguns exames ou questionários mais complexos podem demorar anos para serem refeitos por questões de tempo, disponibilidade de recursos ou até mesmo do paciente. Por isso, em algumas situações, pode ter sido utilizada mais de uma visita para o preenchimento do *dataset*.

Outro ponto importante é que no grupo de pacientes que desenvolveram discinesia durante o estudo, cada paciente começa a apresentar sintomas discinéticos em um momento diferente. Por exemplo, um paciente pode apresentar discinesia na V01, enquanto outro apenas na V12.

Para os primeiros experimentos, que visaram encontrar a melhor configuração possível para os classificadores, foi adotada uma estratégia para diminuir a quantidade de *missing values* e ter os dados imediatamente antes do aparecimento da discinesia, ou seja, dados de visitas imediatamente anteriores à que a discinesia foi detectada pela primeira vez, ou seja, a mais recente possível. A Figura 7 mostra o histograma com a diferença em meses dessa abordagem, a partir da qual, é possível extrair que, usando os dados

do acompanhamento mais recente, o tempo médio em relação ao desenvolvimento da discinesia será de 9 meses com um desvio padrão 6,1 meses.

**Figura 7** – Histograma mostrando o intervalo entre a avaliação do paciente e o início da discinesia para pacientes do grupo PD com discinesia.




**Fonte:** O autor

Para a última etapa do estudo, foi realizada uma exploração temporal a fim de identificar com quanto tempo de antecedência os classificadores, com a informação disponível, são capazes de predizer o surgimento da discinesia. Com isso, avaliamos as limitações e o alcance dos nossos modelos. Conforme esperado, a quantidade de *missing values* aumentou nessa etapa, já que nem todos os exames e avaliações são feitos todos os anos.

A Tabela 1 mostra, em anos, os casos considerados no Experimento 5 deste trabalho para a antecedência da avaliação em relação ao surgimento da discinesia para cada intervalo. Para todos os casos, a variação permitida foi de até doze meses para mais ou para menos. Por exemplo, no Intervalo 2, no qual o tempo de antecedência é de 2 anos, se o paciente foi diagnosticado com discinesia em Janeiro de 2020, os dados usados nessa etapa do experimento serão de Janeiro de 2018 permitindo uma uma variação de até 12 meses, ou seja, de Janeiro de 2017 a Janeiro de 2019, e caso exista mais de uma visita nesse período será escolhida a mais próxima de Janeiro de 2018.

### 3.1.3 Resumo dos dados

O banco de dados PPMI apresenta uma ampla gama de informações coletadas dos pacientes ao longo do tempo. Após extenso estudo dos dados fornecidos pelo banco, as planilhas contendo dados sociodemográficos e da doença, provenientes de avaliações neurológicas, motoras e não motoras foram identificadas e organizadas para o presente

**Tabela 1** – Antecedência das avaliações empregadas neste trabalho, em anos, em relação ao desenvolvimento da discinesia e a variação permitida em meses.

| | Ano |
|---|---|
| Intervalo 1 | 1 ano ± 12 meses |
| Intervalo 2 | 2 anos ± 12 meses |
| Intervalo 3 | 3 anos ± 12 meses |
| Intervalo 4 | 4 anos ± 12 meses |
| Intervalo 5 | 5 anos ± 12 meses |

**Fonte:** O autor

estudo, e estão representadas nas Tabelas 2, 3, 4 e 5, respectivamente. Nas tabelas, estão indicadas as características usadas como entradas dos classificadores.

**Tabela 2** – Informações gerais.

| Feature | Descrição |
|---|---|
| PATNO | Identificador único do paciente do banco de dados. |
| Discinesia | Identifica se tem discinesia ou não, calculado pela MDS-UPDRS parte IV. |
| Gênero | Identifica se é masculino ou feminino. |
| Data do nascimento | Usado para cálculo da idade do paciente. |
| Data do diagnóstico | Usado para cálculo da idade de diagnóstico e tempo com discinesia. |
| Histórico Familiar | Histórico familiar de Parkinson. |
| Educação | Número de anos de educação. |
| LEDD | Quantidade de LEDD que o paciente tomou, pode ser identificado de acordo com o medicamento. |

**Fonte:** O autor

**Tabela 3** – Informações neurológicas.

| Feature | Descrição |
|---|---|
| Caudate | Denervação dopaminérgica nessa região, detectada através de tomografia computadorizada por emissão de fóton para detecção do transportador de dopamina (DAT). |
| Putamen | Denervação dopaminérgica nessa região, detectada através de tomografia computadorizada por emissão de fóton para DAT. |
| CSF $\alpha$-synuclein | Quantificação da proteína alfa-sinucleína no líquido cefalorraquidiano. |

**Fonte:** O autor

**Tabela 4** – Informações motoras.

| Feature | Descrição |
|---|---|
| Modified Schwab + England ADL | Escala de mobilidade para pacientes com mobilidade reduzida. |
| PASE | Escala de atividade física para idosos. |
| MDS-UPDRS II | Experiências motoras do dia a dia do paciente. |
| MDS-UPDRS III | Sinais motores da DP antes e depois administração de medicação. |
| MDS-UPDRS IV | Complicações motoras como: discinesia e flutuações motoras que incluem distonia. |

**Fonte:** O autor

**Tabela 5** – Informações não motoras.

| Feature | Descrição |
|---|---|
| REM | Escala do sono REM. |
| Smell Test | Teste de cheiro da Universidade da Pensilvânia. |
| ESS | Escala de sonolência de Epworth. |
| QUIP | Desordens impulsivas-compulsivas nos portadores da DP. |
| Semantic fluency | Tarefa que avalia fala de determinados grupos semânticos. |
| HVLT | Teste de aprendizagem verbal Hopkins. |
| Letter-Number Sequencing | Sequenciamento de letras-números é uma medida de capacidade de memória de trabalho. |
| SCOPA-AUT | Avaliação de disfunção autonômica da escala de desfechos na DP. |
| SDMT | Teste de Modalidades de Dígitos de Símbolos. |
| STAI-Y | O Inventário de Ansiedade Traço-Estado é uma auto-avaliação que mede a ansiedade do estado emocional em adultos. |
| MoCA | Escala que avalia a memória de curto prazo e de trabalho, habilidades viso espaciais, função executiva, atenção, concentração, linguagem e orientação. |
| GDS | Escala de depressão geriátrica curta. |
| JLOT | Teste de Orientação da Linha (JLOT) é uma medida de percepção e orientação espacial. |
| SCOPA-AUT | Escala de Avaliação de Distúrbios Autonômicos na DP. |
| MDS-UPDRS I | Experiências não motoras do dia a dia do paciente. Dividida em duas partes, onde uma é respondida pelo avaliador e a outra pelo próprio paciente. |

**Fonte:** O autor

### 3.1.4 *Feature engineering*

Essa etapa foi uma das mais importantes do projeto, pois nela foi obtido valores que são derivados de um ou mais dados e que trazem um sentido mensurável para cada informação. Algumas variáveis estão dispostas de forma mais simples, como por exemplo idade, gênero e escolaridade, em que um cálculo com datas ou seleções já fornece a

informação desejada.

Em outros casos, no entanto, foi necessário entender a forma de avaliação feita para calcular um escore final com base nas respostas e nos pesos de cada item. Exemplos desse tipo de situação foram a escala de sono Epworth, a escala de depressão e o teste de memória verbal de curto prazo de Hopkins. Cada uma dessas variáveis tem os dados brutos extraídos do banco de dados, a partir dos quais é feita a soma dos itens de interesse para a geração do escore representativo.

Já no que diz respeito à dose de medicamentos, foi necessário consultar uma tabela de fármacos e cruzar os dados com a data ou o intervalo de datas, a frequência de uso e o tamanho da dose de cada droga. Esses casos tornam-se ainda mais complexos em alguns medicamentos que possuem alguma ação dopaminérgica e deve ser considerada a dose diária equivalente de levodopa (LEDD, do inglês Levodopa Equivalent Daily Dosage), como por exemplo os Inibidores da Catecol-O-Metiltransferase (ICOMT). A LEDD é a dose do medicamento necessária para que os efeitos do mesmo sejam equivalentes aos de 100 mg de levodopa (JULIEN et al., 2021). Cada variável resultante do processamento foi usada como entrada para as análises e será, a partir de agora, referida como *feature* neste documento.

### 3.1.5 Sumarização dos dados

Depois da aplicação dos critérios, 707 pacientes foram selecionados com 49 *features*. Destas, três são categóricas, ou seja, são marcadas com 1 ou 0 caso a característica esteja ou não presente, respectivamente. Um resumo pode ser visto na Tabela 6 que, além do total de pacientes, dos pacientes com mutação genética e dos dados categóricos, mostra a distribuição daqueles que apresentam ou não discinesias.

**Tabela 6** – Sumarização dos dados categóricos.

| | Com discinesia | Sem discinesia | Total |
|---|---|---|---|
| Pacientes no estudo | 241 (34.1%) | 466 (65.9%) | 707 |
| Mutação genética | 84 (30.3%) | 193 (69.7%) | 277 |
| Gênero masculino | 144 (34.7%) | 271 (65.3%) | 415 |
| Gênero feminino | 97 (33.7%) | 191 (66.3%) | 288 |
| Possui histórico familiar de DP | 85 (31.3%) | 187 (68.7%) | 272 |

**Fonte:** O autor

A grande maioria dos dados é contínua, como mostra a Tabela 7 com suas respectivas médias, desvios padrão e valores mínimos e máximos. Com isso, é possível ter uma ideia de se os valores mais comuns ficam próximos do mínimo, máximo ou do centro, e a sua variação.

**Tabela 7** – Sumarização dos dados discretos e contínuos.

| | média | desvio padrão | mínimo | máximo |
|---|---|---|---|---|
| age | 62.25 | 9.98 | 29.00 | 85.00 |
| age_of_pd_diagnosis | 59.80 | 10.07 | 28.00 | 84.00 |
| years_with_pd | 2.46 | 1.86 | 0.00 | 12.00 |
| education | 15.45 | 3.61 | 0.00 | 26.00 |
| jlo_totraw | 11.96 | 2.70 | 0.00 | 15.00 |
| tremor_score | 4.94 | 4.70 | 0.00 | 26.00 |
| hopkins | 0.80 | 0.27 | 0.00 | 2.00 |
| mseadlg | 84.34 | 14.01 | 0.00 | 100.00 |
| number_sequencing | 9.51 | 3.19 | 0.00 | 20.00 |
| pase_score | 138.34 | 92.16 | 0.00 | 504.67 |
| scopa_aut | 13.95 | 7.87 | 0.00 | 48.00 |
| sdm_test | 36.72 | 12.96 | 0.00 | 68.00 |
| alpha_synuclein | 1510.01 | 724.44 | 356.10 | 8405.70 |
| mean_caudate | 1.92 | 0.58 | 0.39 | 3.87 |
| mean_putamen | 0.81 | 0.33 | 0.15 | 2.43 |
| asymmetry_index_caudate | 0.50 | 23.18 | -73.79 | 84.78 |
| asymmetry_index_putamen | -2.15 | 42.07 | -134.07 | 115.15 |
| contralateral_caudate | 2.06 | 0.62 | 0.36 | 4.08 |
| contralateral_putamen | 0.91 | 0.40 | 0.14 | 2.79 |
| ledd | 937.99 | 1572.63 | 0.00 | 12343.75 |
| ledd_levodopa | 514.93 | 671.00 | 0.00 | 7620.00 |
| ledd_dopamine_agonists | 106.17 | 564.16 | 0.00 | 12120.00 |
| ledd_amantadine | 30.83 | 88.14 | 0.00 | 800.00 |
| ledd_maob_inhibitors | 44.17 | 51.43 | 0.00 | 300.00 |
| score_upd_i | 2.51 | 2.85 | 0.00 | 17.00 |
| score_upd_i_pq | 7.24 | 4.25 | 0.00 | 21.00 |
| score_upd_ii | 10.74 | 7.49 | 0.00 | 42.00 |
| score_upd_iii | 17.56 | 17.70 | 0.00 | 86.00 |
| score_upd_iii_a | 6.70 | 12.53 | 0.00 | 68.00 |
| score_upd_i_sum | 9.75 | 6.14 | 0.00 | 36.00 |
| score_upd_sum_total | 38.05 | 23.61 | 0.00 | 146.00 |
| pigd | 0.50 | 0.54 | 0.00 | 3.60 |
| td | 0.45 | 0.41 | 0.00 | 2.36 |
| td_pigd | 1.21 | 1.39 | 0.00 | 10.45 |
| stai_y | 69.28 | 20.44 | 17.00 | 138.00 |
| s_anxiety | 34.47 | 10.82 | 17.00 | 70.00 |
| t_anxiety | 34.81 | 10.84 | 0.00 | 68.00 |
| smell_test | 21.19 | 8.76 | 0.00 | 40.00 |
| epworth_sleepiness_scale | 7.59 | 4.75 | 0.00 | 24.00 |
| semantic_fluency | 42.10 | 15.99 | 0.00 | 87.00 |
| score_quip | 0.44 | 0.87 | 0.00 | 7.00 |
| mcatot | 25.76 | 4.22 | 0.00 | 30.00 |
| score_rem | 4.79 | 3.18 | 0.00 | 13.00 |
| score_gds | 5.82 | 1.81 | 1.00 | 13.00 |

**Fonte:** O autor

## 3.2 PRÉ-PROCESSAMENTO

Na etapa de pré-processamento, foram realizados todos os ajustes para uma entrada de dados consistente no classificador, de forma que não houvesse favorecimento para nenhuma das classes por conta de maiores quantidades ou das *features* por um maior intervalo de variação, entre outras coisas. Os passos foram realizados na ordem descrita abaixo.

### 3.2.1 Normalização

Os dados numéricos foram normalizados de 0 a 1, o que significa dizer que a variação de todos os dados passou a ser de 0 até 1, utilizando o padrão MinMax descrito na Equação (2), onde $x_i$ é o valor a ser normalizado, $max(X)$ é o valor máximo coletado para a *feature* $X$, $min(X)$ é o valor mínimo e $z_i$ é o valor normalizado. Já para os dados categóricos, foram criadas colunas e marcadas com 1 quando possuíssem a característica e 0 quando não.

$$z_i = \frac{x_i - min(X)}{max(X) - min(X)} \tag{2}$$

### 3.2.2 *Missing values*

Inicialmente, foi utilizado um método de exclusão em registros com mais de um terço das *features* com *missing values* com o objetivo de excluir pacientes que abandonaram a pesquisa ou que ainda não possuíam dados suficientes para esse estudo. Para os demais casos, foram utilizados três métodos distintos de preenchimento.

O preenchimento dos *missing values* foi realizado usando duas abordagens. Na primeira, foram utilizados dois métodos: substituição pela média dos valores, que é uma forma muito utilizada na literatura e tem como principal característica não modificar a distribuição estatística dos dados, e o preenchimento com zero, nos casos em que a falta de um valor indica a inexistência daquela *feature* para o paciente em questão, como no caso do LEDD. Na segunda abordagem, a média foi substituída pelo MissForest que emprega o Random Forest para fazer previsões para preenchimento usando como base os registros preenchidos de uma forma mais automática e personalizada para cada caso (STEKHOVEN; BÜHLMANN, 2012).

A Tabela 8 mostra um resumo da quantidade de *missing values* e os métodos utilizados para preenchimento, destacando também o preenchimento com zero que não foi modificado. Essas duas diferentes formas de preenchimento foram utilizadas em experimentos diferentes, como mostra na Tabela 8, o uso da média foi exclusivo do Experimento 1 e o uso do MissForest foi no Experimento 2 e nos demais, já que acabaram seguindo a partir dele, como ilustrado na Figura 6.

**Tabela 8 –** Quantidade de *missing values* e forma de preenchimento.

| *Feature* | *Missing values* | Experimento 1 | Experimento 2 |
|---|---|---|---|
| alpha_synuclein | 220 | média | MissForest |
| td_pigd | 138 | média | MissForest |
| mcatot | 75 | média | MissForest |
| contralateral_caudate | 72 | média | MissForest |
| mean_caudate | 72 | média | MissForest |
| mean_putamen | 72 | média | MissForest |
| asymmetry_index_caudate | 72 | média | MissForest |
| contralateral_putamen | 72 | média | MissForest |
| asymmetry_index_putamen | 72 | média | MissForest |
| pase_score | 56 | média | MissForest |
| td | 41 | média | MissForest |
| pigd | 41 | média | MissForest |
| ledd_comt_inhibitors | 30 | zero | zero |
| ledd_dopamine_agonists | 30 | zero | zero |
| ledd_levodopa | 30 | zero | zero |
| ledd | 30 | zero | zero |
| ledd_amantadine | 30 | zero | zero |
| ledd_maob_inhibitors | 30 | zero | zero |
| years_with_pd | 4 | média | MissForest |
| age_of_pd_diagnosis | 4 | média | MissForest |
| age | 4 | média | MissForest |
| smell_test | 3 | média | MissForest |
| sdm_test | 2 | média | MissForest |
| jlo_totraw | 2 | média | MissForest |
| family_history_of_pd | 2 | média | MissForest |
| number_sequencing | 1 | média | MissForest |
| epworth_sleepiness_scale | 1 | média | MissForest |

**Fonte:** O autor

### 3.2.3 Importância das *features*

Para calcular a importância de cada uma das *features*, foi necessário treinar previamente um classificador e em seguida consultar qual a importância que cada *features* teve no resultado daquele treinamento. Vários classificadores podem ser usados para este fim e cada classificador, hiperparâmetros escolhidos e até mesmo as sementes dos números aleatórios podem interferir nesse resultado, já que cada combinação levará a um treinamento diferente e por consequência o conhecimento extraído será um pouco diferente também. Em todos os experimentos realizados neste trabalho, foi utilizado o Random Forest como classificador de referência.

### 3.2.4 Análise de componentes principais

Nesse projeto, a PCA foi utilizada em conjunto com o conceito de importância de *features*, no qual aquelas menos importantes, obtidas do treinamento, tiveram suas principais características extraídas e depois colocadas de volta com a dimensionalidade reduzida. Esse passo foi realizado com o objetivo de deixar o treinamento mais simples para o classificador, restringindo algumas *features* e obtendo uma quantidade menor de componentes em que a variação é maior. Com isso, buscou-se reduzir o risco do conhecido problema chamado de maldição da dimensionalidade.

### 3.2.5 Balanceamento de dados

Como pode ser visto na Tabela 6, existe uma diferença significativa entre a quantidade de registros com e sem discinesia. Esse é um problema comum que pode levar o modelo a enviesar seus resultados à classe com maior número de registros.

Uma das formas mais comuns de resolver esse problema é excluir alguns dados da classe que possui mais registros de forma que haja um balanceamento entre os dados, ou seja, 50% para cada classe. Apesar de este ter sido o método utilizado nesse trabalho, entende-se que, para um *dataset* com um pequeno número de amostras, essa não é a solução ideal.

### 3.2.6 Divisão do *dataset*

Para esse trabalho, foi feita uma divisão no *dataset* em treinamento, validação e teste. A partir dessa divisão, os conjuntos ficaram com 60%, 20% e 20% dos dados utilizados, respectivamente. Nesses três conjuntos criados os dados foram escolhidos aleatoriamente e sem repetição, então os pacientes usados nos testes, por exemplo, não aparecem no conjunto de treinamento ou de validação. Além disso, os dados estão balanceados em cada um dos três conjuntos, como descrito na Seção 3.2.5.

## 3.3 TREINAMENTO

Para o desenvolvimento desse trabalho, foi utilizada a linguagem Python na versão 3.7 e a biblioteca de aprendizado de máquina scikit-learn na versão v0.20 que possui os modelos Multilayer Perceptron (MLP), Support Vector Machine (SVM), Árvore de Decisão, Random Forest, AdaBoost e Regressão Logística, além do RandomizedSearchCV e de todas as métricas de desempenho necessárias para os experimentos.

### 3.3.1 Busca de hiperparâmetros

A busca de hiperparâmetros foi realizada usando o RandomizedSearchCV (implementação do scikit-learn para o *grid search* com limite de busca) e limitada para no

máximo 1000 diferentes combinações de parâmetros, o que, pela quantidade inseridas, na maioria dos classificadores é mais do que o suficiente para completar todas as combinações do *grid search*. Nesse projeto, foram utilizados todos os hiperparâmetros e combinações que a biblioteca permitiu, com exceção daqueles que a combinação de parâmetros é infinita, como no caso do MLP, em que a quantidade de camadas e de neurônios pode variar indefinidamente.

### 3.3.2 Modelos utilizados

Essa seção irá descrever os parâmetros utilizados na busca pelo RandomizedSearchCV para cada um dos classificadores.

**AdaBoost:** a busca de parâmetros variou a quantidade de estimadores entre 10 e 100 com incremento de 10 e com taxa de aprendizado de 0,1, 0,5, 1 e 1,5.

**Árvore de Decisão:** a busca foi entre os parâmetros de qualidade do corte da árvore, "gini" ou "entropy". A estratégia para escolher a divisão em cada nó foi "best" ou "random". O número de *features* consideradas para o melhor corte foram a "auto", "sqrt", "log2" e a "quantidade total". Por ser apenas uma árvore, a altura não foi limitada.

**Multilayer Perceptron:** a busca de parâmetros variou entre uma e duas camadas escondidas, cada uma com de 1 a 100 nerônios, usando as funções de ativação "identity", "logistic", "tanh" e "relu" e com a otimização de pesos "lbfgs", "sgd" e "adam".

**Support Vector Machine:** a busca de parâmetros variou a regularização C como uma progressão geométrica de 10 variando entre 0.001 e 10 e as funções *kernel* utilizadas foram "rbf", "poly" e "sigmoid".

**Regressão Logística:** a busca de parâmetros variou a quantidade máxima de iterações entre 50, 100 e 150, o parâmetro de regularização C como uma progressão geométrica de 10 variando de 0.001 a 10 e a otimização "newton-cg", "lbfgs", "liblinear", "sag' e "saga".

**Random Forest:** esse é um caso de classificador que pode variar seus parâmetros infinitamente. Então, a quantidade de árvores foi limitada em 100, 200, 300, 500, 750 e 1000 e a profundidade máxima de cada árvore foi de 80, 90, 100 até 110. Além disso, os parâmetros informados no item Árvore de Decisão também foram usados.

### 3.3.3 Medidas de desempenho

Durante o treinamento, para a escolha do melhor modelo, foi utilizada a acurácia como medida de desempenho. Além disso, para cada uma das trinta execuções, um

gráfico boxplot foi montado com a acurácia a fim de demonstrar a sua variação para os classificadores. Adicionalmente, para o melhor modelo, também será mostrada a curva ROC juntamente com o cálculo da AUC, para ajudar na interpretação e comparação entre os classificadores.

# 4 RESULTADOS

Esse trabalho foi realizado em diversas etapas de treinamentos e melhorias dos modelos. Inicialmente, com os dados recém extraídos da base de dados, foi realizado o Experimento 1 que será utilizado como linha de base para comparação com as melhorias aplicadas nos próximos passos. A segunda etapa, o Experimento 2, foi desenvolvido com o objetivo de melhorar a forma de preenchimento dos *missing values* usando o MissForest. Em seguida, a terceira etapa, o Experimento 3, que consistiu em utilizar o conceito de *feature importance* com aplicação da PCA. É importante não confundir esta linha de base de comparação entre classificadores com a linha de base definida na base PPMI. A menos que informado o contrário, a partir deste ponto, quando for citado o termo linha de base, o mesmo refere-se à comparação entre abordagens.

As etapas seguintes foram realizadas para testar as limitações e o alcance do método, considerando a preocupação em tornar essa ferramenta viável para o uso clínico. Um importante passo, o Experimento 4, diz respeito a reduzir o número e identificar quais *features* são as mais informativas ou relevantes para os modelos. O outro passo, o Experimento 5, considerando a realidade clínica, diz respeito a identificar o instante ótimo em que as avaliações devem ser realizadas para identificar aqueles pacientes que estão em maior risco de desenvolver discinesia no futuro. Portanto, essas etapas finais do estudo consistiu em identificar quais são as *features* necessárias para a predição da discinesia, bem como qual a antecedência máxima que o nosso método permite predizê-la.

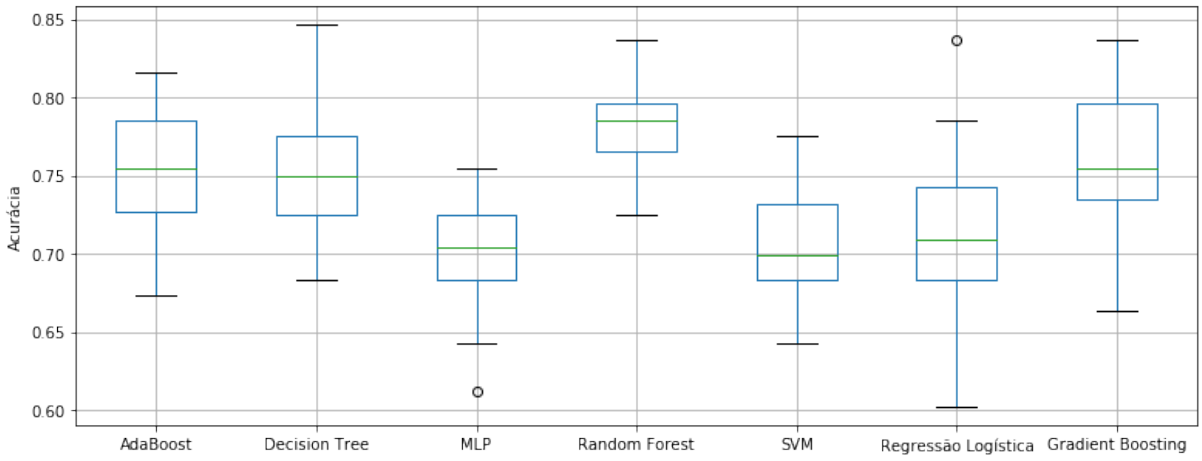
## 4.1 BUSCA DO MELHOR CLASSIFICADOR

Inicialmente, foi realizado um treinamento com todos os dados gerados na mineração com o preenchimento de *missing values* usando a média e o zero como mostrado no "Experimento 1" da Tabela 8, para todos os modelos discutidos o Seção 3.3.2. Essa primeira fase foi considerada a linha de base para comparação com as melhorias propostas.

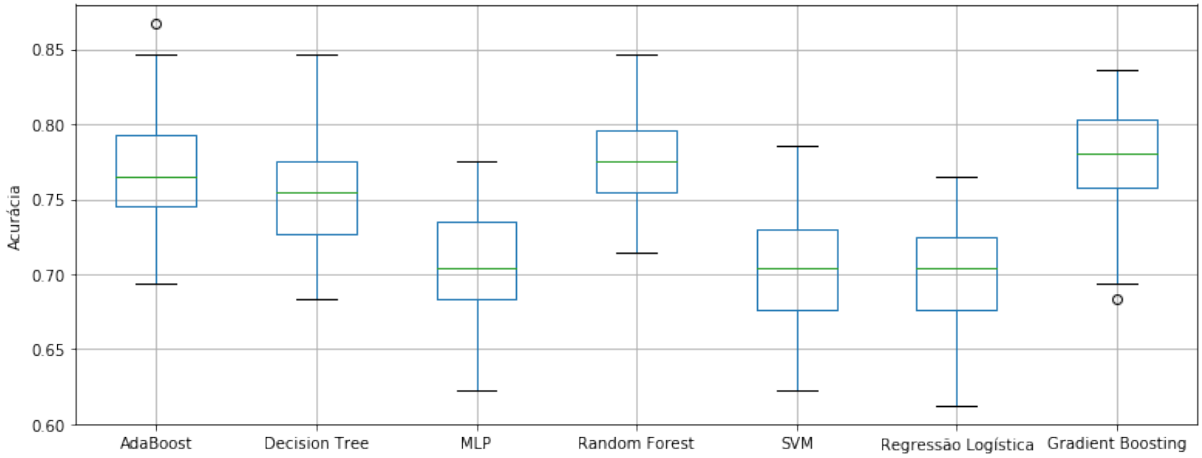
Foram feitos trinta treinamentos diferentes e a Figura 8a mostra a variação da acurácia obtida nos dados de teste para as execuções. Inicialmente, nota-se que o Random Forest mostrou um desempenho mais consistente que os demais, além de menor variação nos resultados.

A Figura 13 disponível no Anexo A mostra a curva ROC e a área sob a curva para as melhores execuções de cada modelo. Apesar da performance baixa da Regressão Logística, quando medida pela acurácia, a sua melhor execução, e somente essa, foi tão boa quanto a do Random Forest e do Gradiente Boosting, resultando numa área sob a curva próxima dos demais modelos.

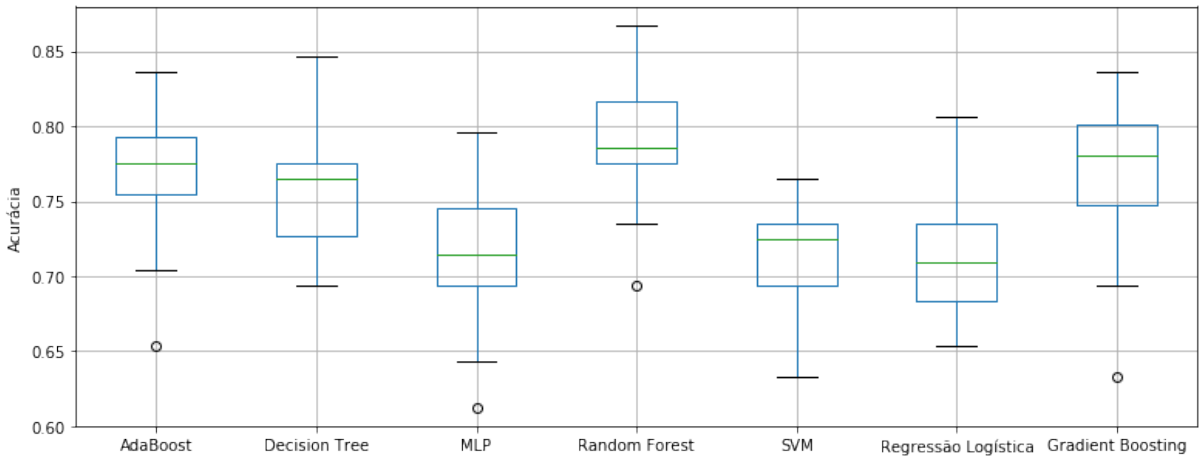
**Figura 8** – Boxplot da acurácia em cada classificador executado trinta vezes.




**(a)** Resultados da linha de base




**(b)** Resultados com o MissForest




**(c)** Resultados com o PCA

**Fonte:** O autor

A primeira melhoria aplicada foi a substituição do método de preenchimento de *missing values* pelo MissForest, conforme a coluna Experimento 2 na Tabela 8. Algumas delas permaneceram usando o valor zero padrão por conta de características do próprio dado.

Após a entrada dos dados faltantes usando o método o MissForest, os modelos foram novamente executados. A Figura 8b mostra as acurácias das trinta novas execuções para cada um dos modelos. Nessa figura, pode ser observada uma leve melhoria dos resultados de alguns modelos como o AdaBoost e o Gradient Boosting. Observando as novas curva ROC de cada um dos classificadores, ilustradas na Figura 14 disponíveis no Anexo A, pode ser visto que, aparentemente, o Gradient Boosting teve uma piora no desempenho, enquanto os demais aparentam ter melhoras ou permaneceram com a mesma AUC.

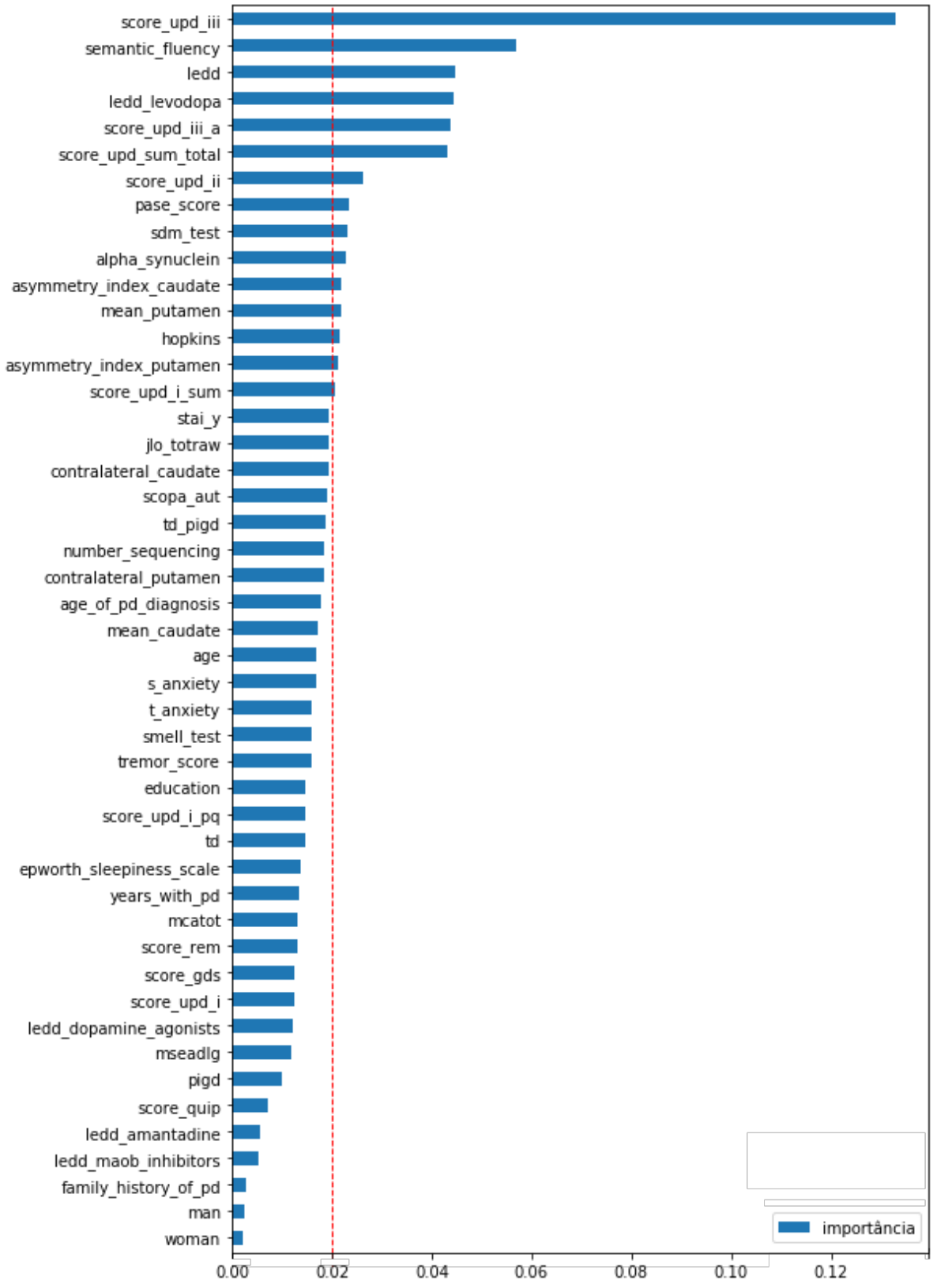
O trabalho de mineração de dados deu origem inicialmente a 48 *features* além do alvo *dyskinesia*, que mostra pacientes que desenvolveram discinesia. A Figura 9 mostra a importância de cada *feature* extraída do classificador Random Forest e pode ser visto a importância que o classificador deu as *features* ligadas a UPDRS e ao LEDD, que aparecem em seis das sete primeiras.

A fim de refinar as *features* incluídas na execução dos modelos, realizamos um corte no valor $0,02$, a partir do qual as *features* que têm maior importância serão usadas diretamente, como mostrado na Figura 9 onde a linha vermelha representa o ponto de corte. As demais *features* passaram por um processo de análise por componentes principais (PCA) que resultou em dez *features*. Com isso, esperava-se que o classificador ficasse mais simples já que a quantidade de entradas diminuiu. Essa etapa de pré-processamento aparentemente gerou melhorias principalmente no Random Forest, no MLP e na Regressão Logística, como pode ser visto na Figura 8c.

Analisando a curva ROC e a AUC, na Figura 15 disponíveis no Anexo A, apenas os classificadores Regressão Logística e o Gradiente Boosting tiveram melhorias a partir do ranqueamento das *features* e aplicação do PCA, enquanto os demais permaneceram iguais ou tiveram um leve decréscimo, lembrando que esse gráfico é gerado apenas com o melhor caso.

De acordo com os resultados apresentados até o momento e complementado na compilação disponível na Tabela 9, o Random Forest mostrou melhores resultados, de forma mais consistente e com menores variações da curva ROC e da acurácia. Para simplificação das próximas etapas e de suas análises, vamos desconsiderar os outros classificadores e acompanhar apenas os resultados obtidos pelo Random Forest.

**Figura 9** – Cálculo de importância das *features* usando o classificador Random Forest.




**Fonte:** O autor

**Tabela 9** – AUC dos classificadores para as diferentes abordagens.

| Classificador | Linha de base | MissForest | PCA |
|---|---|---|---|
| AdaBoost | 85% | 92% | 86% |
| Árvore de Decisão | 88% | 88% | 88% |
| MLP | 79% | 83% | 83% |
| Random Forest | 90% | 91% | 90% |
| SVM | 80% | 82% | 79% |
| Regressão Logística | 83% | 80% | 87% |
| Gradient Boosting | 90% | 85% | 89% |

**Fonte:** O autor

## 4.2 MUDANÇA DE *FEATURES*

Antes da realização dos próximos experimentos, algumas mudanças foram aplicadas no *dataset*. A primeira delas foi a inclusão do tempo de uso dos medicamentos. Então, para cada um dos medicamentos, foi calculado a diferença de tempo em meses que o paciente começou a fazer uso até a data de referência. Outra modificação aplicada foi adicionar LEDD dos Inibidores da COMT, que como consequência aumentou um pouco o LEDD total.

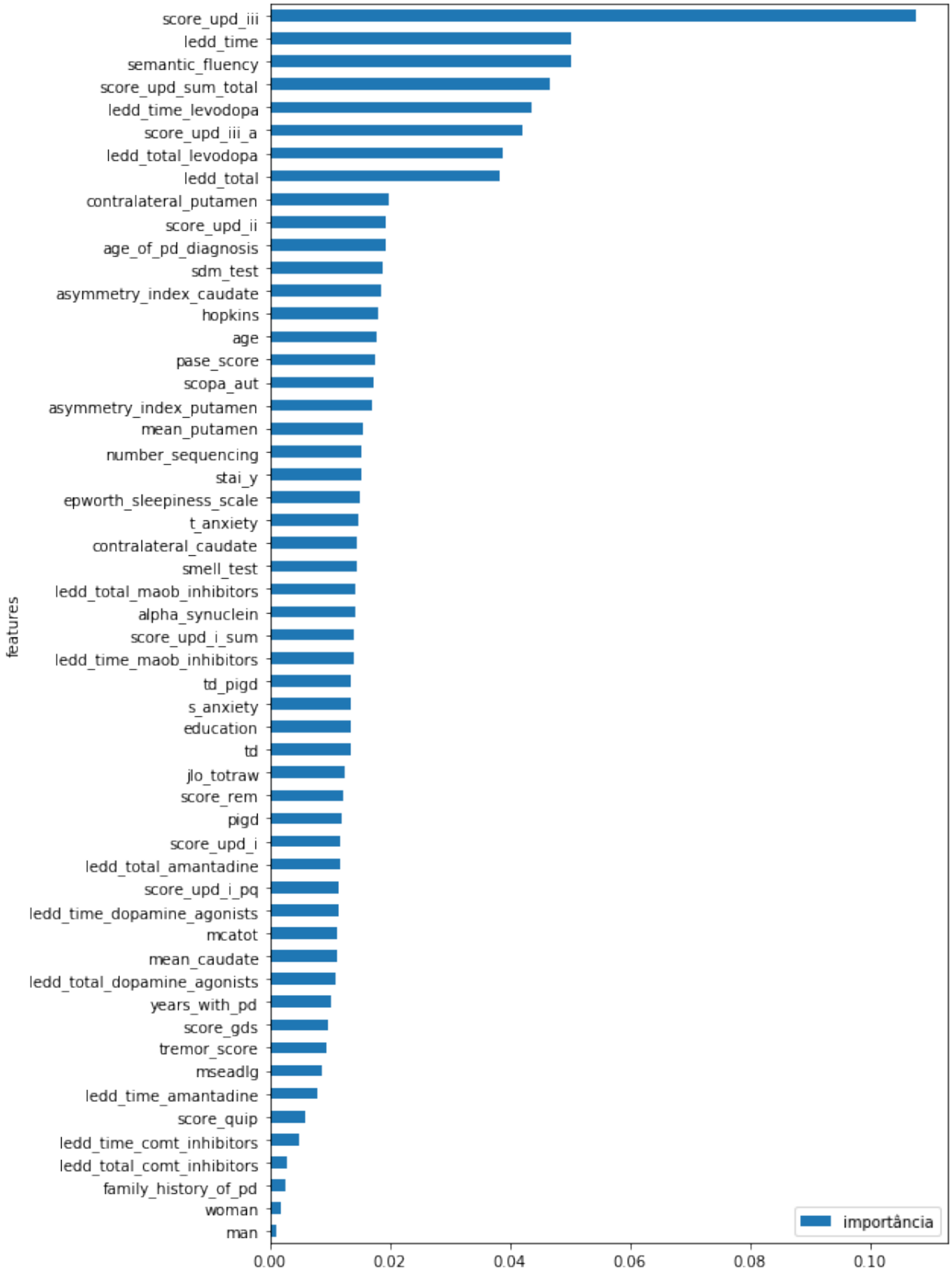
Com as mudanças aplicadas, os resultados gerais de acurácia não foram modificados, porém a ordem da importância das *features* sofreu algumas modificações. Como é importante para o cálculo das próximas etapas, a Figura 10 mostra a nova ordem que foi utilizada na etapa de verificação da quantidade mínima de recursos necessários.

## 4.3 QUANTIDADE MÍNIMA DE *FEATURES*

Pensando na simplificação da aplicação do modelo, o Experimento 4 visa saber se realmente todas as 54 *features* coletadas são necessárias para manter uma acurácia nos valores que foram obtidos. Para isso, serão utilizados os resultados gerados na análise de *feature importance*, disponíveis na Figura 10, e variar a quantidade de *features* iniciando com um e indo até todas disponíveis.

Os resultados dessa variação usando o Random Forest está disponível na Figura 11, onde pode ser visto que a acurácia máxima foi obtida com 16 *features* e ainda com uma mediana superior à maioria dos seus pares e com uma variação menor que o gráfico com oito, por exemplo, que foi outro caso com bons resultados. Os próximos resultados não superaram o obtido com 16 *features*, mas mantiveram o limite superior da acurácia um pouco inferior, com pequenas variações. Essa análise é especialmente importante pois indica que, ao invés de coletar 54 diferentes dados dos pacientes, muitos deles obtidos através de exames neurológicos ou aplicação de escalas psicológicas, apenas 16 primeiras

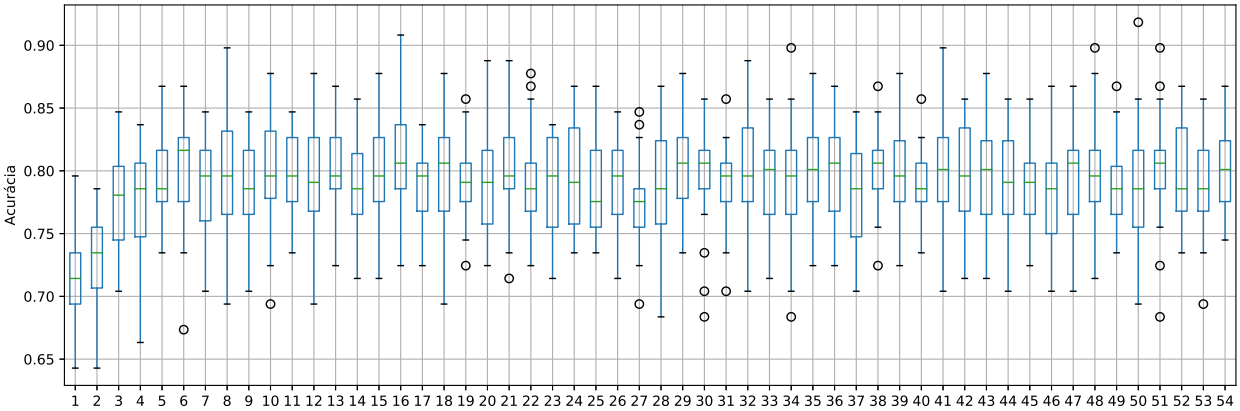
**Figura 10** – Cálculo de importância das *features* atualizado usando o classificador Random Forest.




**Fonte:** O autor

*features* da Figura 10 são suficientes para obter um resultado satisfatório, sendo que vários delas estão concentradas na UPDRS e no LEDD, melhorando ainda mais o cenário, já que pode ser visto como menos exames ou variáveis coletadas.

**Figura 11** – Boxplot da variação da quantidade de *features* usando o classificador Random Forest.




**Fonte:** O autor

## 4.4 DETERMINANDO A MÁXIMA ANTECEDÊNCIA DE TEMPO PARA PREVISÃO

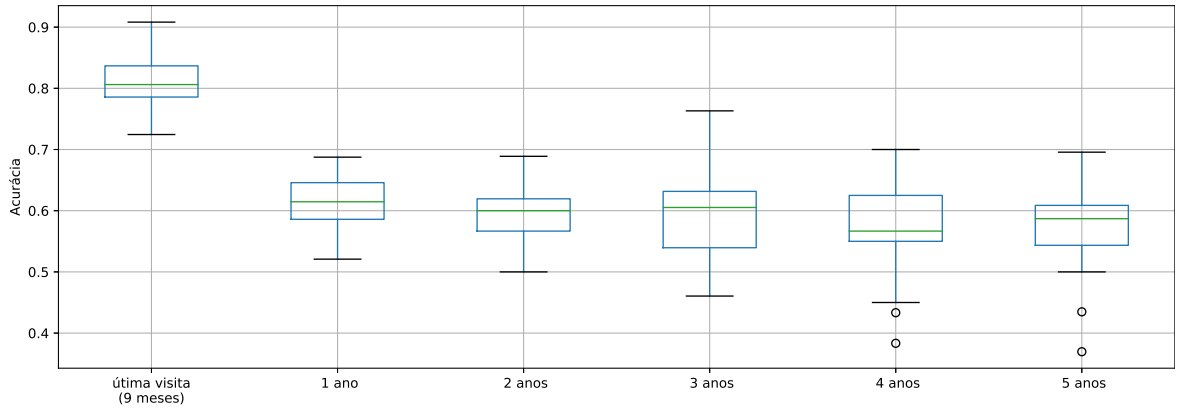
Esta análise, o Experimento 5, teve o objetivo de investigar com que antecedência, em anos, é possível ter uma previsão razoável do desenvolvimento da discinesia utilizando o nosso método. Em um contexto clínico, é mais interessante trabalhar de forma mais antecipada ou preventiva. Logo, quanto maior o tempo de antecedência com que a previsão é realizada, mais mudanças podem ser feitas no tratamento, para evitar a discinesia.

Até então, o classificador havia sido alimentado com dados obtidos da última coleta realizada antes do desenvolvimento da discinesia. O novo experimento terá limitação temporal dos dados utilizados conforme apresentado na Tabela 1, discutida na Seção 3.1.2. Foi avaliada a capacidade de predição de discinesia pelo Random Forest quando foram considerados dados obtidos com 1 a 5 anos de antecedência à ocorrência da discinesia. Pensando em incrementar o experimento anterior, aqui foi mantida a limitação de 16 *features*, usando as mesmas do experimento anterior, que são aquelas de maior relevância mostradas na Figura 10.

Esse experimento teve um desempenho inferior dos que vínhamos trabalhando, como pode ser visto na Figura 12, na qual o primeiro boxplot refere-se ao desempenho de 16 *features* da Seção 4.3. Em relação aos intervalos do experimento atual, temos que os de um e dois anos tiveram um valor máximo de acurácia um pouco menor que o intervalo de

três anos, o que demonstra ser um resultado diferente do esperado, já que, quanto mais próximo temporalmente da discinesia, mais aparentes espera-se que sejam os sintomas. Porém, olhando também para a mediana e para a variação do resultado, o gráfico de um ano de antecedência mostra mais consistência com um limite inferior mais distante de 50%. Usando a mesma abordagem das etapas anteriores, os valores de AUC calculados no ponto máximo dos boxplots da Figura 12 estão disponíveis na Tabela 10, onde podemos notar a significativa queda em relação aos experimentos anteriores.

**Figura 12** – Boxplot da acurácia mudando o distanciamento em anos da discinesia.




**Fonte:** O autor

**Tabela 10** – AUC dos classificadores para os diferentes intervalos de tempos.

| 1 ano | 2 anos | 3 anos | 4 anos | 5 anos |
|---|---|---|---|---|
| $75,9\%$ | $72,6\%$ | $79,9\%$ | $70,6\%$ | $66,5\%$ |

**Fonte:** O autor

Uma possível explicação para essa piora nos resultados é a quantidade de *missing values*, que ficou mais frequente com essa restrição temporal. A Tabela 11 resume essa diferença entre as abordagens, onde a segunda coluna mostra a quantidade de registros que entram na análise depois do filtro, que restringe quem tem mais de um terço dos dados preenchidos, e a terceira mostra a quantidade de *missing values* que são preenchidos pelo MissForest, já retirando os casos do LEDD, que sua ausência é entendida real, ou seja, o paciente não faz uso da medicação, e contando somente para as 16 *features* selecionadas. Na primeira visita disponível para o intervalo de um ano, a quantidade de *missing values* aumenta muito, quase um terço, e nos outros anos a quantidade não aumenta mais porque os registros acabam sendo excluídos, chegando a excluir mais da metade (51%) dos pacientes no intervalo de cinco anos.

**Tabela 11** – Quantidade de registros e de *missing values* de acordo com a variação de tempo.

| | Registros | *Missing values* |
|---|---|---|
| Última visita | 704 (-11.4%) | 201 |
| 1 ano | 704 (-11.4%) | 263 |
| 2 anos | 645 (-18.9%) | 228 |
| 3 anos | 555 (-30.2%) | 208 |
| 4 anos | 468 (-41.1%) | 186 |
| 5 anos | 388 (-51.2%) | 176 |

**Fonte:** O autor

Com essa configuração experimental, os dados acabam ficando sobrepostos, ou seja, a variação de 1 ano mais 12 meses coincide com o de 2 anos menos 12 meses. Isso pode influenciar nos resultados, deixando até um pouco semelhante. Porém, foi usado preferencialmente dados mais próximos do centro, ou seja, com menor variação até o ano alvo. O ideal teria sido, obviamente, trabalhar com menores intervalos de tempo, no entanto, isso tornou-se inviável devido ao aumento enorme de *missing values*.

# 5 DISCUSSÃO

Durante a realização desse trabalho, várias etapas de um projeto de mineração de dados foram realizadas como entendimento do problema, entendimento dos dados e a preparação dos dados. Mesmo que essas etapas não tenham ficado em destaque no resultado final, elas foram de extrema importância e influenciam diretamente os resultados, além de terem sido as que mais exigiram tempo.

A partir dos experimentos desenvolvidos, foi encontrado um classificador capaz de identificar pacientes que desenvolveram discinesia com acurácia de 90,8% e AUC ROC de 93,8%, que representa uma taxa de acerto muito boa para esse tipo de previsão, mesmo considerando que foi obtida com um tempo médio de nove meses de antecedência. Com base nessa previsão, médicos e pacientes têm mais alguns pontos de análise para definir estratégias para retardar o surgimento da discinesia. Em conjunto com essa classificação, também foram identificadas as principais variáveis que influenciaram essa previsão. Esse é outro achado interessante que pode guiar pesquisas e no futuro até fazer parte da decisão para tratamento de pacientes.

Para chegar no melhor classificador apresentado foram considerados os resultados de acurácia e AUC ROC para os melhores casos. Porém, para uma possível aplicação, é importante observar também qual a mediana e a variação dos resultados, pois com esses pontos é possível ter uma boa ideia da possível resposta do modelo a situações não observadas até o momento.

Nas outras etapas do projeto, foram investigadas as limitações e o alcance desse método em relação ao menor uso de recursos e o maior tempo de antecedência para a previsão. Através das análises realizadas foi encontrado que, com mais de 16 *features* a acurácia deixa de apresentar melhoras. Então, usar somente essas 16 representa uma economia bem significativa de recursos, já que a quantidade inicial era 48 *features* para serem coletadas, armazenadas e processadas. Na análise temporal, os dados mostram que, quando consideramos o tempo mais curto, a previsão é muito superior. Porém, caso o objetivo seja realizar a predição com maior antecedência, a previsão com três anos teve os melhores resultados, com uma variação menor e maior mediana. Entretanto, é importante mencionar uma vez mais que as baixas acurácias desses modelos podem ter sido agravadas pelo aumento na quantidade de *missing values* nesse período e pelo fato das variáveis terem apresentado sobreposição nos anos próximos.

As *features* mais importantes apontadas pelo classificador estão relacionadas à UPDRS III e II, ao tempo e à dose de LEDD, à fluência semântica e à denervação do Putâmen, conforme pode ser visto em detalhe na Figura 10. Esses resultados são

similares aos encontrados por Eusebi et al. (2018), que também usou um classificador para determinar quais seriam os fatores de risco para o desenvolvimento de discinesias. As principais diferenças em relação ao nosso estudo dizem respeito à inclusão da variável risco genético, e a importância do gênero. Enquanto para Eusebi et al. (2018) ser do gênero feminino foi considerado um fator de risco para discinesia, no nosso estudo, a variável gênero foi uma das menos importantes. Essas diferenças podem ser explicadas pelas atualizações no próprio banco de dados, já que o banco é alimentado continuamente com novas informações, e nosso acesso mais recente aconteceu em 2020, dois anos após a publicação do artigo citado, e também por diferenças em como as *features* foram processadas, a inclusão do grupo com mutação genética e o uso de um classificador diferente. No estudo "Discinesias induzidas por levodopa são precedidas por ansiedade e agravamento dos prejuízos motores em pacientes da Doença de Parkinson" realizado por nos (Dias, C.; Brys, I.; e Leal, D. em 2020), ainda não publicado, também encontramos resultados semelhantes. Naquele estudo, usamos estatísticas multivariadas para comparar, antes mesmo do início da terapia dopaminérgica, pacientes que desenvolveram discinesias com aqueles que não desenvolveram. Outro estudo que encontrou resultados semelhantes aos nossos foi o de Nicoletti et al. (2016), que reuniu dados de 485 voluntários diagnosticados com a DP, dos quais 128 desenvolveram discinesias. Os autores encontraram que a duração da doença, o estágio Hoehn-Yahr, o escore do paciente na UPDRS, o gênero feminino e o tempo de uso de medicação com ação dopaminérgica foram fatores associados ao desenvolvimento de discinesias.

Este trabalho traz como principal resultado um classificador capaz de atuar como um sistema de suporte à decisão de tratamento de pacientes da DP, que pode ser usado em conjunto com outros estudos afim de contribuir para mudanças no tratamento dos pacientes, como diminuição da dose de levodopa. Um trabalho que pode vir a contribuir com isso é o realizado por Liu et al. (2020), no qual dados de 403 pacientes da DP foram analisados com o objetivo de encontrar qual seria a dose máxima de levodopa para o paciente não desenvolver discinesia. Os autores fizeram uso de Estatística, Árvore de Decisão e Regressão Linear e da curva ROC e concluiram que 400 mg por dia parece ser a dose máxima de levodopa possível a fim minimizar o desenvolvimento de discinesias. Os autores também trouxeram abordagens interessantes para o uso de variáveis como a dose de levodopa pelo peso do paciente, para ter uma ideia da quantidade do medicamento que está sendo tomado por cada quilo, além de usar o próprio peso do paciente como entrada para os modelos.

Apesar da grande quantidade inicial de *features* utilizadas no presente estudo, é importante mencionar que ainda outras poderiam ser aplicadas para buscar resultados melhores. Em um estudo publicado recentemente, Severson et al. (2021) usaram aprendizado de máquina para encontrar estágios de progressão da DP. Segundo os autores, um diferencial do estudo foi adotar uma estratégia capaz de fazer o modelo considerar a complexidade da medicação, o que, segundo eles, foi um diferencial em relação a trabalhos anteriores.

Severson et al. (2021) dividiram a progressão da doença em oito estágios, do mais simples para o mais severo, e encontraram que pacientes que iniciaram a doença já no quinto estágio tiveram uma evolução mais rápida. Como os oito estágios descritos por Severson et al. (2021) relacionam-se com o desenvolvimento de discinesias e ainda uma questão a ser respondida, e estudos futuros podem ser realizados no sentido de compreender essa relação. Além do estágio da progressão da doença, outros trabalhos já citados neste capítulo também apresentaram *features* que foram importantes nos respectivos trabalhos e poderiam contribuir para melhores resultados do nosso classificador, como o estágio Hoehn-Yah usado por Nicoletti et al. (2016) e o risco genético usado por Eusebi et al. (2018).

Mesmo com a enorme qualidade do banco de dados PPMI e da sua quantidade de dados, que é considerada grande quando comparada com outras bases da Doença de Parkinson, o volume de dados pode ser considerado ainda limitado, considerando a complexidade do problema. Além disso, a base de dados apresenta outras limitacoes, como *missing values* e desbalanceamento de dados. Para diminuir o problema dos *missing values*, foi usada a estratégia de buscar o último dado coletado, mesmo que essa coleta tenha acontecido há menos tempo. Este problema foi resolvido apenas parcialmente, tendo permanecido um alto número de *missing values*. Na etapa em que a quantidade de anos é restringida, podemos ver como o problema de *missing values* afeta o desempenho. Isso se dá pela forma de trabalho e coleta de dados, pois nem sempre é viável fazer a coleta de todas as variáveis, já que elas estão dispostas em 141 tabelas e a maioria delas tem mais de uma variável para ser inserida.

O problema da quantidade e do desbalanceamento serão resolvidos à medida que o PPMI lançar novas versões com mais pacientes, o que fará provavelmente a quantidade de pacientes com discinesia também aumentar. O *oversampling* foi uma estratégia estudada para acabar com o desbalanceamento, mas houve dúvidas se essa injeção de dados acabaria influenciando positivamente o modelo e dando uma visão não realista do problema. Além disso, outra abordagem poderia ter sido a utilização de outra medida de desempenho como a AUC ROC no treinamento, o que poderia ter sido feito no início do estudo. Entretanto, devido à grande quantidade de experimentos já implementados, a mudança de medida de desempenho iria atrapalhar a comparação das várias abordagens e dos experimentos realizados anteriormente.

Nos últimos cinco anos, identificar pacientes da DP em maior risco de desenvolver discinesia parece ter sido uma importante questão de pesquisa, e estudos têm sido publicados com o objetivo de fornecer informações sobre a doença e mostrar o que pode estar mais relacionado com a discinesia. Este trabalho também mostrou quais variáveis foram mais importantes na tomada de decisão do classificador, mas foi além e propôs um sistema de suporte à decisão. Além de indicar quais informações têm mais relevância, o presente

estudo fornece uma previsão antecipada da provável ocorrência de discinesia para apoio à decisão clínica que possa mudar doses ou outros fatores para que o paciente tenha uma menor chance de desenvolver a discinesia. Esse foi um trabalho pioneiro na sua formulação do problema, na montagem dos dados e na metodologia usando aprendizado de máquina. Resultados como os nossos abrem possibilidades para novos estudos clínicos que visam prevenir ou retardar o início da discinesia em pacientes da DP.

Esse projeto exigiu conhecimentos em áreas que normalmente são estudadas separadamente como Ciência da Computação, Farmácia e Psicologia, além de áreas com naturezas mais interdisciplinares como a Neurociência e a Ciência de Dados. Desde o início, esse projeto mostrou ser um bom exemplo de trabalho interdisciplinar e a importância de trabalhar dessa forma. Durante toda a fase de mineração, de proposição de experimentos e de análise dos resultados exigiu o trabalho em conjunto de profissionais de diferentes áreas.

# 6 CONCLUSÃO

A experimentação foi dividida em etapas de melhoria de desempenho e testes de limitações e do alcance. Como outros trabalhos evolutivos, primeiro foi construído o primeiro protótipo com os dados recém minerados. Em seguida, foi feita uma melhoria na fase de preenchimento de *missing values*, depois foi feito o uso da PCA para diminuir a dimensionalidade e tentar usar os pontos mais relevantes da grande quantidade de *features* extraídas. Nas fases finais, pensando na aplicabilidade da nossa tecnologia à realidade clínica, investigamos qual seria o número mínimo de *features* para ter um modelo sem queda de desempenho e qual o intervalo de tempo máximo em que a predição poderia ser realizada. Essas últimas análises são especialmente importantes para que a predição da discinesia seja feita de forma otimizada e com o máximo de antecedência, de forma que haja mais tempo para o desenvolvimento de estratégias de prevenção ou que visem retardar o início dos sintomas discinéticos.

Na implementação das diferentes fases dos experimentos, o melhor resultado encontrado foi obtido usando-se apenas 16 *features* com o classificador Random Forest, no qual a acurácia foi de $90,8\%$, com sua mediana $79,6\%$, e sua AUC ROC de $93,8\%$, com sua mediana $87,0\%$. O uso de todo o conjunto de *features* mineradas foi importante para encontrar o melhor modelo, mas mais que isso, para encontrar quais as mais importantes e para montar uma entrada mais simples. Esse resultado é muito animador para o projeto, visto que uma abordagem adotada com o intuito de simplificar o uso em contexto clínico, foi capaz de gerar resultado equivalente ao inicial.

## 6.1 TRABALHOS FUTUROS

O PPMI fornece muitas informações que não foram exploradas neste trabalho, algumas por fugir do tema ou do objetivo da pesquisa e outras por complexidade na obtenção e tratamento como, por exemplo, exames de imagens e material genético. Alguns estudos podem ser feitos buscando entender a evolução de sintomas que não são tão explorados pela literatura, como os sintomas não motores. Além desses pontos que podem render bons frutos para essa linha de pesquisa, também existe a possibilidade de mudar a estratégia de uso dos dados, pensando em novas formas de pré-processamento e aplicação de novos modelos.

# REFERÊNCIAS


ABAS, M. A. H. et al. Agarwood oil quality classification using support vector classifier and grid search cross validation hyperparameter tuning. **Int. J**, v. 8, 2020. Citado na página 21.

AGRAWAL, P. **SVM in R for Data Classification using e1071 Package**. 2020. TechVidvan. Disponível em: <https://techvidvan.com/tutorials/svm-in-r/>. Acesso em: 24 jun 2021. Citado na página 19.

AROWOLO, M. O.; ADEBIYI, M. O.; ADEBIYI, A. A. An efficient pca ensemble learning approach for prediction of rna-seq malaria vector gene expression data classification. **Int. J. Eng. Res. Technol.**, v. 13, n. 1, p. 163–169, 2020. Citado na página 21.

BALESTRINO, R.; SCHAPIRA, A. H. Parkinson disease. **European journal of neurology**, Wiley Online Library, v. 27, n. 1, p. 27–42, 2020. Citado 4 vezes nas páginas 13, 22, 23 e 24.

BALLABIO, D.; GRISONI, F.; TODESCHINI, R. Multivariate comparison of classification performance measures. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 174, p. 33–44, 2018. Citado na página 21.

BRUCE, P.; BRUCE, A. **Practical Statistics for Data Scientists**. [S.l.]: O'Reilly Media, 2017. Citado na página 17.

BRYS, I. et al. Neurophysiological effects in cortico-basal ganglia-thalamic circuits of antidyskinetic treatment with 5-ht1a receptor biased agonists. **Experimental neurology**, Elsevier, v. 302, p. 155–168, 2018. Citado na página 24.

CENCI, M. A. et al. Dyskinesia matters. **Movement Disorders**, Wiley Online Library, v. 35, n. 3, p. 392–396, 2020. Citado na página 23.

CHEN, W. et al. Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. **Journal of Hydrology**, Elsevier, v. 583, p. 124602, 2020. Citado na página 19.

DANGETI, P. **Statistics for machine learning**. [S.l.]: Packt Publishing Ltd, 2017. Citado 3 vezes nas páginas 16, 17 e 18.

DAS, K.; BEHERA, R. N. A survey on machine learning: concept, algorithms and applications. **International Journal of Innovative Research in Computer and Communication Engineering**, v. 5, n. 2, p. 1301–1309, 2017. Citado 2 vezes nas páginas 15 e 19.

DORSEY, E. R. et al. Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. **Neurology**, AAN Enterprises, v. 68, n. 5, p. 384–386, 2007. Citado na página 13.

EUSEBI, P. et al. Risk factors of levodopa-induced dyskinesia in parkinson's disease: results from the ppmi cohort. **npj Parkinson's Disease**, Nature Publishing Group, v. 4, n. 1, p. 1–6, 2018. Citado 4 vezes nas páginas 13, 14, 48 e 49.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 21.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. [S.l.]: Alta Books, 2019. Citado 3 vezes nas páginas 16, 21 e 22.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, Springer, p. 1–11, 2021. Citado na página 16.

JULIEN, C. et al. The clinical meaning of levodopa equivalent daily dose in parkinson's disease. **Fundamental & Clinical Pharmacology**, Wiley Online Library, v. 35, n. 3, p. 620–630, 2021. Citado na página 31.

KINGE, D.; GAIKWAD, S. Survey on data mining techniques for disease prediction. **International Research Journal of Engineering and Technology (IRJET)**, v. 5, n. 01, p. 630–636, 2018. Citado 2 vezes nas páginas 18 e 19.

KLUG, M. et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. **Journal of general internal medicine**, Springer, v. 35, n. 1, p. 220–227, 2020. Citado na página 20.

KUHN, W.; MÜLLER, T. Amantadine for treating parkinson's disease. **NeuroPsychopharmacotherapy**, Springer, p. 1–6, 2020. Citado na página 24.

LIASHCHYNSKYI, P. **Creating of neural network using JavaScript in 7 minutes!** 2019. Dev. Disponível em: <https://dev.to/liashchynskyi/creating-of-neural-network-using-javascript-in-7minutes-o21>. Acesso em: 25 jun 2021. Citado na página 18.

LIU, G. et al. Risk thresholds of levodopa dose for dyskinesia in chinese patients with parkinson's disease: a pilot study. **Neurological Sciences**, Springer, v. 41, n. 1, p. 111–118, 2020. Citado na página 48.

MAREK, K. et al. The parkinson progression marker initiative (ppmi). **Progress in neurobiology**, Elsevier, v. 95, n. 4, p. 629–635, 2011. Citado na página 13.

NICOLETTI, A. et al. Clinical phenotype and risk of levodopa-induced dyskinesia in parkinson's disease. **Journal of neurology**, Springer, v. 263, n. 5, p. 888–894, 2016. Citado 2 vezes nas páginas 48 e 49.

PAI, V. V.; PAI, R. B. Artificial intelligence in dermatology and healthcare: An overview. **Indian Journal of Dermatology, Venereology and Leprology**, Scientific Scholar, p. 1–11, 2021. Citado na página 20.

ROSCHER, R. et al. Explainable machine learning for scientific insights and discoveries. **IEEE Access**, IEEE, v. 8, p. 42200–42216, 2020. Citado na página 16.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. Citado na página 19.

SEVERSON, K. A. et al. Discovery of parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. **The Lancet Digital Health**, Elsevier, 2021. Citado 2 vezes nas páginas 48 e 49.

SOLANA-LAVALLE, G.; ROSAS-ROMERO, R. Classification of ppmi mri scans with voxel-based morphometry and machine learning to assist in the diagnosis of parkinson's disease. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 198, p. 105793, 2021. Citado 2 vezes nas páginas 14 e 15.

STEKHOVEN, D. J.; BÜHLMANN, P. Missforest—non-parametric missing value imputation for mixed-type data. **Bioinformatics**, Oxford University Press, v. 28, n. 1, p. 112–118, 2012. Citado na página 33.

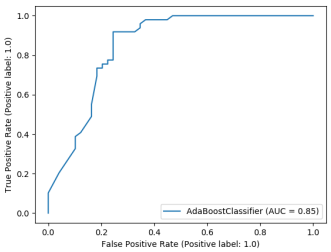
TIBCO. **What is a Random Forest?** 2021. Disponível em: <https://www.tibco.com/reference-center/what-is-a-random-forest>. Citado na página 20.

UNPINGCO, J. **Python for probability, statistics, and machine learning**. [S.l.]: Springer, 2016. v. 1. Citado 2 vezes nas páginas 18 e 19.

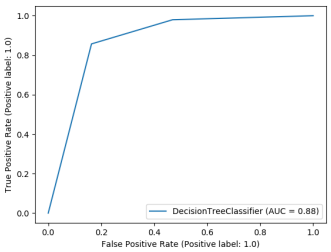
ZESIEWICZ, T. A.; SULLIVAN, K. L.; HAUSER, R. A. Levodopa-induced dyskinesia in parkinson's disease: epidemiology, etiology, and treatment. **Current Neurology and Neuroscience Reports**, Springer, v. 7, n. 4, p. 302–310, 2007. Citado na página 13.

# ANEXO A – CURVA ROC DAS MELHORES EXECUÇÕES DE CADA UM DOS MODELOS

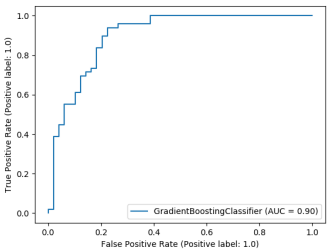
**Figura 13** – Curva ROC das melhores execuções de cada um dos modelos na linha de base.



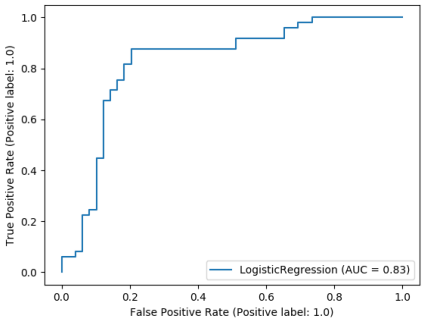

**(a)** AdaBoost



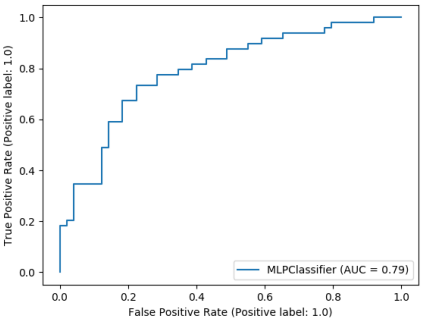

**(b)** Árvore de Decisão



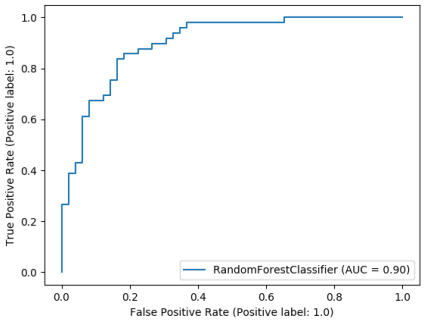

**(c)** Gradient Boosting



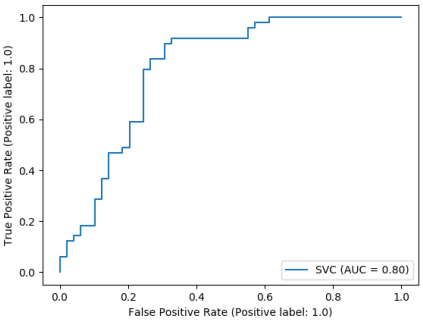

**(d)** Regressão Logística




**(e)** MLP



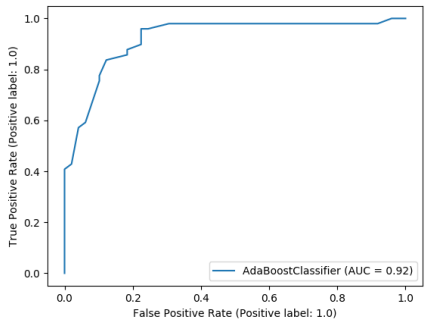

**(f)** Random Forest



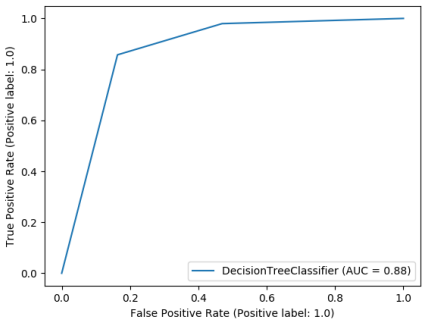

**(g)** SVM

**Fonte:** O autor

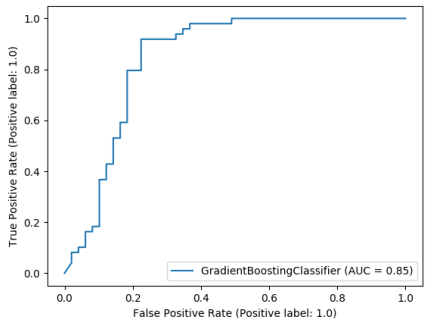
**Figura 14** – Curva ROC das melhores execuções de cada um dos modelos depois do tratamento de *missing values*.



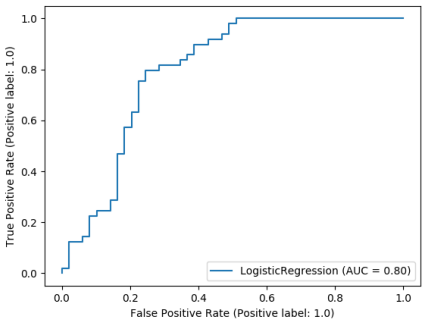

**(a)** AdaBoost



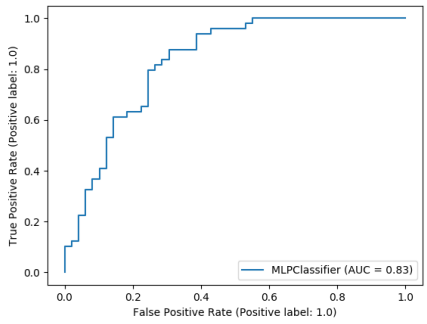

**(b)** Árvore de Decisão



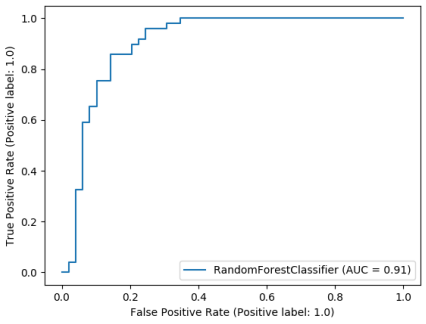

**(c)** Gradient Boosting



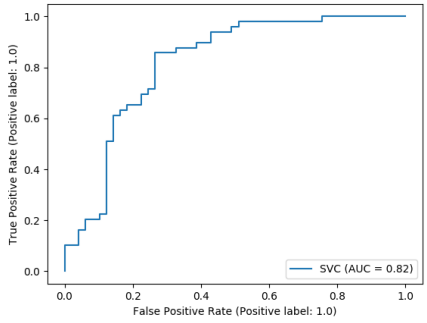

**(d)** Regressão Logística




**(e)** MLP



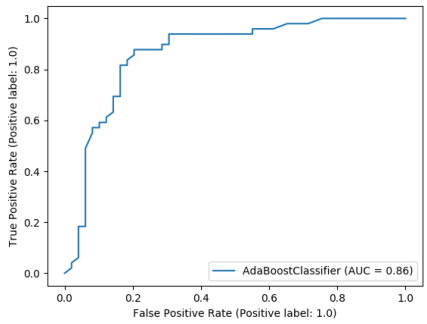

**(f)** Random Forest



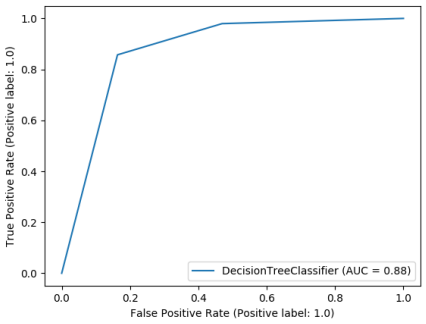

**(g)** SVM

**Fonte:** O autor

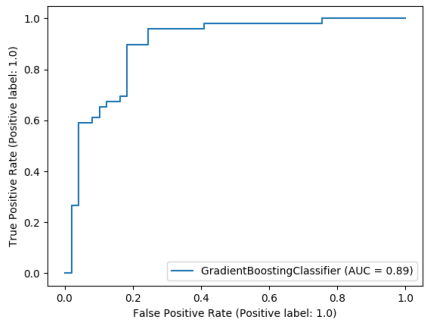
**Figura 15** – Curva ROC das melhores execuções de cada um dos modelos depois de colocar o PCA.



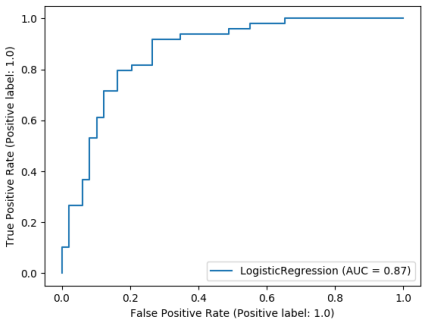

**(a)** AdaBoost



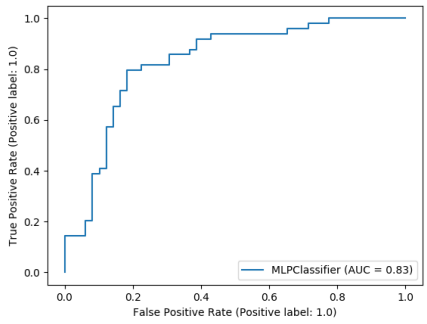

**(b)** Árvore de Decisão



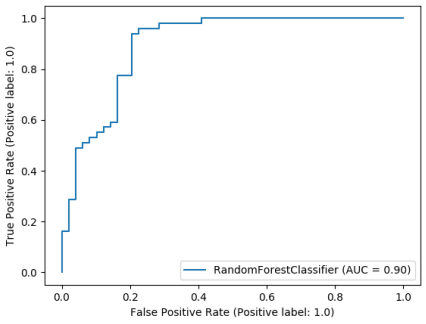

**(c)** Gradient Boosting



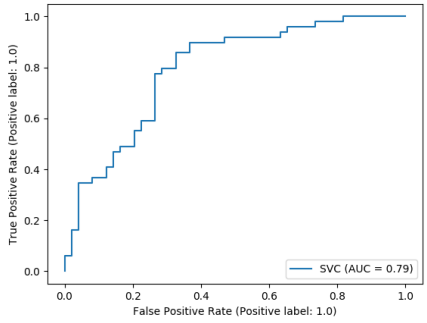

**(d)** Regressão Logística




**(e)** MLP




**(f)** Random Forest




**(g)** SVM

**Fonte:** O autor