**UNIVERSIDADE FEDERAL DE ALFENAS**
**INSTITUTO DE CIÊNCIAS EXATAS**
**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

*Eugênio Ferreira Cabral*

# ANÁLISE DA IMPUTAÇÃO DE DADOS EM ÁRVORES DE CLASSIFICAÇÃO BINÁRIA

Alfenas, 03 de Agosto de 2017.

**UNIVERSIDADE FEDERAL DE ALFENAS**
**INSTITUTO DE CIÊNCIAS EXATAS**
**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

# ANÁLISE DA IMPUTAÇÃO DE DADOS EM ÁRVORES DE CLASSIFICAÇÃO BINÁRIA

*Eugênio Ferreira Cabral*

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Alfenas como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Humberto César Brandão de Oliveira

Alfenas, 03 de Agosto de 2017.

*Eugênio Ferreira Cabral*

# ANÁLISE DA IMPUTAÇÃO DE DADOS EM ÁRVORES DE CLASSIFICAÇÃO BINÁRIA

A Banca examinadora abaixo-assinada aprova a monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação pela Universidade Federal de Alfenas.

---

**Prof. Dr. Humberto César Brandão de Oliveira**

**Universidade Federal de Alfenas**

---

**Prof. Dr. Luiz Eduardo da Silva**

**Universidade Federal de Alfenas**

---

**Prof. Dr. Ricardo Menezes Salgado**

**Universidade Federal de Alfenas**

Alfenas, 03 de Agosto de 2017.

Dedico este trabalho aos meus pais.

# AGRADECIMENTO

Agradeço primeiramente aos meus pais, Luiz Antônio Cabral e Selma de Fátima Ferreira, por todo apoio e cuidados à mim prestado, além de serem fontes de inspiração e exemplo durante toda minha caminhada.

Aos amigos feitos durante todos estes anos de graduação, pela troca de experiências e aprendizados, os quais foram fundamentais para meu desenvolvimento. Em especial aos meus amigos, Guilherme de Oliveira Santos, Talysson Oliveira Cassiano e Vinícius Ferreira da Silva.

Ao meu orientador Humberto César Brandão de Oliveira, pela confiança, apoio e paciência durante todos os anos em que trabalhamos juntos. Agradeço por cada lição aprendida, por cada erro cometido, e por cada conquista. Seus ensinamentos, os quais sou muito grato, foram de grande valia em minha formação.

À Universidade Federal de Alfenas (UNIFAL) e ao Laboratório de Pesquisa e Desenvolvimento (LP&D) pelas oportunidades fornecidas ao longo dos anos.

Aos fomentos providos pela PROBIC/UNIFAL e PIBITI/CNPq para o desenvolvimento de minhas Iniciações Cientificas e ao CNPq pelo programa Ciência Sem Fronteiras onde enriqueci ainda mais minha formação.

Aos professores, e todos aqueles que contribuíram para minha formação e me auxiliaram de alguma forma a concluir esta jornada.

A todos, muito obrigado!

"Some people worry that artificial intelligence will make us feel inferior, but then, anybody in his right mind should have an inferiority complex every time he looks at a flower."

Alan Kay

# RESUMO

A classificação de padrões é um subtópico da área de aprendizagem de máquina, no qual o objetivo é detectar e reconhecer padrões através do uso de algoritmos de aprendizado. Porém, para efetuar a classificação, antes é necessário avaliar a integridade dos dados. Nessa avaliação, há casos nos quais são possíveis de detectar a presença de dados faltantes, o que pode ocasionar em classificações erradas. Portanto, para contornar esse problema, surge como alternativa a necessidade de estimar os valores faltantes, com a intenção de manter a integridade da informação, e minimizar ou eliminar a possível perda de precisão na classificação de padrões. O processo que estima os valores faltantes é conhecido como imputação de dados, onde é possível encontrar diversos algoritmos que podem ser usados neste processo. Porém considero que pouco se sabe sobre o impacto desses algoritmos em determinados tipos de classificadores. Com base nisso, este trabalho busca compreender melhor o impacto dos diferentes algoritmos de imputação, considerando classificadores baseados em árvore de decisão. Nos experimentos realizados durante a execução deste trabalho, concluí que embora exista uma perda gradual de precisão nos classificadores, independente do classificador e do imputador utilizado, identifiquei que o classificador J48 é mais sensível à perda de dados que o *Random Forest*. Além disso, observei que sem o uso de imputadores, os classificadores produzem resultados similares ou melhores em relação à pontuação original enquanto o percentual de dados faltantes está abaixo da metade. Entretanto, quando o percentual ultrapassa a metade dos dados, a utilização de técnicas de imputação pode reduzir a perda de pontuação dos classificadores.

**Palavras-Chave:** aprendizado de máquina, árvores de decisão, classificação binária, imputação de dados.

# ABSTRACT

Pattern classification is a subtopic of the machine learning discipline, which aims to discover emerging patterns through learning algorithms. However, to perform the classification, it is necessary to test for missing data first, because there are cases where the presence of missing data can lead to misclassifications. In order to maintain the information integrity and reduce the potential score loss in the classification, a possible alternative to overcome the missing data issue, is to estimate the missing values. The process that estimates the missing values is known as data imputation, and there are several algorithms that can be used for this task. However, I consider that little is known about the influence of imputation algorithms on certain types of classifiers. Thus, the aim of this experiment is to understand better the influence of imputation algorithms in tree-based classifiers. In the conducted experiments, I have observed that, regardless of the classifier and the imputation algorithm used, there is a gradual score loss produced by the classifiers as the missing percentage grows, and the classifier J48 is more sensitive to data loss than Random Forest. In addition, when the percentage of missing data is below 50%, without the use of imputation, the classifiers produce similar results compared to the original score. However, when the percentage of missing data exceeds 50%, the usage of imputation algorithms can reduce the score loss produced by the classifiers.

**Keywords:** machine learning, decision trees, binary classification, data imputation

# LISTA DE FIGURAS

# LISTA DE ABREVIAÇÕES

| | |
|---|---|
| MAR | *Missing at Random* |
| MCAR | *Missing Completely at Random* |
| MNAR | *Missing not at Random* |
| AUC | *Area Under the Curve* |
| ROC | *Receiver Operating Characteristic* |
| VP | Verdadeiros Positivos |
| FP | Falsos Positivos |
| KNN | *k-Nearest Neighbors* |
| MI | *Multiple Imputation* |
| MICE | *Multiple Imputation by Chained Equation* |
| MLE | *Maximum Likelihood Estimation* |
| EM | *Expectation Maximization* |

# SUMÁRIO

# 1
# Introdução

*Este capítulo apresenta a motivação e os objetivos deste trabalho. Na Seção 1.1 é descrito a justificativa e motivação do trabalho. Na Seção 0 é apresentado a problematização do tema. Nas subseções 1.3.1 e 1.3.2 são apresentados os objetivos gerais e específicos, respectivamente. E por fim, a Seção 1.4 contém a organização da monografia.*

## 1.1 Justificativa e Motivação

Nos últimos anos houve um crescente aumento na quantidade de dados produzidos por empresas e usuários ao redor do mundo (Liu *et al.,* 1997). Para que todo esse volume de dados seja interpretado apropriadamente com a finalidade de agregar valor, vejo a necessidade do desenvolvimento de técnicas de análise de dados, especialmente algoritmos de aprendizado de máquina. Entretanto, os dados coletados nem sempre estão estruturados da maneira adequada para seu tratamento, ou até mesmo dados faltantes podem estar presentes.

Segundo Gelman e Hill (2006), a presença de dados faltantes em conjuntos de dados pode comprometer a análise ou até mesmo a classificação destes dados, e consequentemente produzir resultados com erros significantes. Uma abordagem que busca solucionar este problema, é conhecida como imputação de dados, que consiste em substituir os valores faltantes por valores não nulos (Gelman e Hill, 2006). Essa abordagem pode produzir diferentes resultados dependendo do tipo de estratégia adotada. Entretanto, os estudos que buscam entender o impacto da utilização de imputadores na classificação de dados, ao meu ver, analisam de forma abrangente e induzem à conclusões genéricas. Em termos práticos, pode levar a uma utilização arbitrária de imputadores, dado que não se tem um entendimento mais específico de seu comportamento. Portanto, em vista de melhor aplicar as técnicas de imputação, este trabalho tem o foco em compreender os impactos dos imputadores na classificação de dados.

## 1.2 Problematização

Considerando que a classificação de dados pode solucionar uma variedade de problemas do mundo real, e que a presença de dados faltantes pode comprometer a qualidade das soluções, acredito ser importante compreender melhor os impactos atrelados à utilização da imputação de dados. Uma vez que existe uma influência direta entre o dado estimado através da imputação de dados e a classificação, considero que a compreensão dessa influência neste contexto seja necessária. Dessa forma, este trabalho se propõe a compreender as limitações e vantagens do tratamento de dados faltantes e seus impactos na classificação de dados, com o intuito de responder a questão: "Existe algum método de imputação que favoreça melhores resultados na classificação de dados?" .

## 1.3 Objetivos

### 1.3.1 Gerais

Este trabalho busca analisar a influência de diferentes algoritmos de imputação na classificação de dados. Dessa forma, o objetivo principal é fornecer conclusões que condicionem à um tratamento mais adequado nos dados faltantes em problemas de classificação de padrões.

### 1.3.2 Específicos

Os objetivos específicos deste trabalho buscam responder as seguintes perguntas:

- Existe algum algoritmo de imputação que favoreça melhores resultados na classificação de dados?
- Os algoritmos de classificação conseguem tratar dados faltantes sem auxílio de algoritmos de imputação?

- Todos algoritmos de classificação tem a mesma sensibilidade à dados faltantes?

## 1.4 Organização da Monografia

O Capítulo 2 introduz o conceito de dados faltantes, apresentando sua definição e os tipos considerados neste trabalho, bem como exemplos para ilustrar os conceitos. Além disso, descreve a definição da imputação de dados de acordo com autores usados nesta monografia, e os diferentes tipos de abordagens utilizadas para realizar a imputação de dados.

O Capítulo 3 introduz o conceito fundamental para este trabalho conhecido como classificação de dados. Além de descrever o funcionamento básico de árvores de decisão, abordagem adotada para as classificações realizadas neste trabalho.

O Capítulo 4 apresenta a revisão bibliográfica do trabalho, contendo os trabalhos e abordagens relacionadas à este experimento em termos de metodologia e ferramentas utilizadas.

O Capítulo 5 apresenta uma visão geral da metodologia adotada neste trabalho, descrevendo em detalhes os passos realizados para a execução do experimento.

O Capítulo 6 contém os resultados obtidos através do experimento, bem como os devidos comentários em relação as análises conduzidas.

O Capítulo 7 apresenta as conclusões do autor, e os trabalhos futuros sugeridos para a continuidade do estudo em questão.

# 2
# Imputação e dados faltantes

*Este capítulo apresenta os conceitos de dados faltantes e imputação de dados. Na Seção 2.1 descreve conceitualmente os dados faltantes e suas implicações. A Seção 2.2 define e exemplifica os tipos de dados faltantes existentes. Na Seção 2.3 é definido a imputação de dados. E por fim, a Seção 2.4 introduz os tipos de imputação de dados e detalha cada um deles nas respectivas subseções.*

## 2.1 Definição de dados faltantes

O conceito de dados faltantes é parte de um conceito mais geral, conhecido como *coarsened data*, no qual incluem números que são agrupados, agregados, arredondados, omitidos, ou truncados, resultando em uma perda parcial de informação (Heitjan, 1991). O conceito mais específico de dados faltantes é definido pela omissão de uma ou de várias informações referentes ao conjunto de dados estudado. Segundo Gelman e Hill (2006), a ocorrência de dados faltantes pode impactar negativamente nas conclusões retiradas dos dados em estudo. Esse tipo de problema pode acontecer por uma falha humana, técnica ou a informação pode não estar disponível durante a coleta. Sendo tópico de pesquisa em áreas do conhecimento que trabalham diretamente com coleta de dados, como por exemplo, economia, meteorologia, sociologia, ciência social e política.

Existem vários indicativos sobre a falta de dados, como por exemplo, "não se aplica", "recusou à informar", "ilegibilidade" entre outros indicadores que podem justificar a inexistência de um dado. Portanto, antes de aplicar algum método para manipular os dados faltantes, é necessário considerar se realmente existe um valor "verdadeiro" para o dado faltante. Para tomar conhecimento de que um valor "verdadeiro" de fato existe, é necessário categorizar o mesmo em algum tipo de dado faltante. Como critério fundamental para distinguir esses tipos, é necessário observar se há uma relação de dependência de dados entre as variáveis em estudo. Na Seção 2.2, é apresentado as definições segundo Allison (2002), de cada tipo de dado faltante e seus respectivos exemplos.

## 2.2 Tipos de dados faltantes

Com o objetivo de compreender os dados faltantes, alguns autores como Alisson (2002) sugerem uma definição objetiva considerando a dependência ou não entre variáveis dos dados faltantes. É importante ressaltar que, devido à natureza subjetiva do problema, não há uma metodologia para garantir a interpretação do tipo de dado faltante, portanto o tipo de dado faltante é assumido, e não determinado.

### 2.2.1 Dado faltante aleatório

Também conhecido como *Missing at Random* (MAR), o dado faltante aleatório é quando a falta de um dado referente à uma variável Y está diretamente relacionada à outra variável X, e não depende de valores de Y. Um nome mais apropriado poderia ser "dados condicionalmente faltantes", pois depende de dados conhecidos e existentes no conjunto de dados (Allison, 2002). Por exemplo, pessoas entrevistadas de determinadas áreas podem ser menos prováveis de fornecer informação sobre seu salário, ou seja o entrevistado pode ou não informar o seu salário (variável Y) dependendo da área em que atua (variável X).

### 2.2.2 Dado faltante completamente aleatório

O tipo conhecido como *Missing Completely at Random* (MCAR) é descrito como um caso especifico do MAR, na qual os valores faltantes na variável Y, não dependem nem de Y nem de outra variável X (Allison, 2002). Por exemplo, quando um entrevistado não pode finalizar o formulário de pesquisa por desistência, ou seja, o dado faltante não tem nenhuma relação com as variáveis do formulário. Embora não seja possível determinar com total certeza à qual tipo o dado faltante pertence, Little (1988) propõe um teste para que se possa ou não assumir MCAR.

### 2.2.3 Dado faltante não aleatório

Existem casos também conhecidos como *Missing not at Random* (MNAR), onde os valores faltantes de uma variável Y dependem da própria variável Y (Allison, 2002). Por exemplo, em uma dada pesquisa, entrevistados com um alto salário podem ser menos prováveis de informar seus próprios salários, ou seja, a falta do dado é dependente da própria variável.

## 2.3 Definição de imputação de dados

A imputação de dados é o processo de substituição de dados faltantes por valores não nulos (Gelman e Hill, 2006). O estudo de técnicas de imputação é motivado pelos resultados controversos gerados através de análises que ignoram as implicações dos dados faltantes. Uma vez que há dados faltantes em um conjunto de dados, um dos primeiros passos é analisar como será o tratamento desse tipo de caso. Na literatura existe uma vasta quantidade de abordagens para contornar o problema da falta de dados, em alguns casos essas abordagens fazem um bom trabalho, em outros, podem eliminar propriedades essenciais do conjunto de dados afetando em maior ou menor grau nos resultados encontrados na análise (Gelman e Hill, 2006). Portanto, a imputação de dados busca preservar informações através de valores estimados usando informações disponíveis. Uma vez que os valores são imputados, os dados podem ser analisados usando técnicas padrões. É importante notar que, para a maioria dos métodos de imputação é assumido que os dados faltantes são MCAR (o dado faltante não depende de nenhuma variável observada), e caso está suposição seja equivocada as análises podem produzir resultados distorcidos (Gelman e Hill, 2006).

## 2.4 Tipos de imputação de dados

Existem várias formas de se aplicar a imputação dos dados, elas são desenvolvidas utilizando diferentes estratégias. Essas estratégias podem ser divididas em grupos

baseando-se em seu funcionamento. Nesta subseção, será apresentado as classificações gerais dos métodos de imputação de acordo com a definição de Gelman e Hill (2006).

### 2.4.1 Descarte de dados

A forma mais simples de tratar dados faltantes, são casos onde as instâncias com dados faltantes são eliminadas do conjunto de dados. Junto à simplicidade desta abordagem, existem algumas implicações com seu uso. A primeira, é a potencial redução do tamanho amostral, o que pode eliminar a representatividade da população estudada e potencialmente elevar o desvio padrão (Gelman e Hill, 2006). Além disso, ao remover informações, alguns testes estatísticos podem se mostrar mais fracos, dificultando assim a busca por sinais (Gelman e Hill, 2006).

### 2.4.2 Imputação de variável única

Essa abordagem engloba vários métodos comuns de imputação de dados, que ao invés de excluir variáveis ou dados faltantes, busca substituir o dado faltante por valores mais apropriados. Uma abordagem comum, é a imputação pela média ou mediana, a qual consiste em estimar o valor a ser imputado baseado na média ou mediana de suas respectivas variáveis. Outras abordagens também conhecidas deste grupo, são a imputação baseada em regras e valores copiados de instâncias anteriores. Como esta abordagem busca estimar os valores nulos baseado em uma única variável, isso pode contribuir para o aumento do *bias* no conjunto de dados (Gelman e Hill, 2006).

### 2.4.3 Imputação baseada em modelo

A imputação baseada em modelo busca considerar um número maior de informações do conjunto de dados para estimar os valores nulos através de algoritmos mais elaborados. Esta abordagem pode ser exemplificada por técnicas de imputação múltipla (*MI*), estimativa por máxima verossimilhança (*MLE*), ou modelos estatísticos. Um dos problemas mais comuns com essa abordagem é o alto custo para encontrar parâmetros adequados para a imputação, em contrapartida, geralmente produz melhores resultados e seu uso é frequentemente sugerido por diferentes autores (Gelman e Hill, 2006).

# 3
# Classificação de dados e árvores de decisão

*Este capítulo introduz os conceitos de classificação de dados e árvores de decisão. Na Seção 3.1 é introduzido de forma conceitual a classificação de dados. A Seção 3.1.1 descreve a forma de avaliação da classificação de dados adotada para este trabalho. Na Seção 3.2 descreve de forma introdutória o funcionamento de árvores de decisão.*

## 3.1 Classificação de dados

Um dos tipos de problemas mais comuns dentro da Inteligência Artificial, é a classificação de dados (Liu *et al.*, 1997). Esse tipo de tarefa tem sido realizado pelos seres humanos desde os primórdios para garantir sua sobrevivência, distinguindo amigos de inimigos, plantas venenosas ou medicinais. Como descrito por Pellegrin (1986), o estudo dessa área se inicia no século IV com Aristóteles, com o intuito de dividir os organismos conhecidos em plantas ou animas, e então subdividindo-os em plantas pequenas, médias e grandes, e animais em terrestres, aquáticos ou voadores. Entretanto, esse sistema de classificação rudimentar não foi suficiente para classificar os organismos encontrados na natureza com grande precisão, mas foi o primeiro passo para o desenvolvimento do ramo da ciência conhecido como taxonomia.

Desde então, a classificação de dados tem se tornado uma abordagem comum para atender diferentes necessidades em diferentes áreas do conhecimento. Atualmente, existe uma variedade de algoritmos computacionais para classificação de dados (Aggarwal, 2014). Esses algoritmos podem ser divididos em dois grandes grupos quanto ao tipo de aprendizado, o aprendizado supervisionado e aprendizado não supervisionado (Russell e Norvig, 2005). No aprendizado supervisionado, seu conceito consiste em aprender a partir de dados rotulados com a resposta correta, como exemplo deste tipo temos algoritmos como redes neurais artificiais, árvores de decisão,

classificadores lineares e redes bayesianas. Por outro lado, o aprendizado não supervisionado consiste em aprender sem nenhuma informação prévia sobre os dados (Russell e Norvig, 2005). Para este grupo, algoritmos como redes neurais artificiais, regras por associação e algoritmos de agrupamento, são bem conhecidos na literatura.

Independentemente do tipo de aprendizado utilizado para resolver o problema, existem tipos de problemas classificação quanto à seu objetivo, dentre eles temos a classificação binária, multiclasse e multirótulo (Aggarwal, 2014). Na classificação binária o objetivo é classificar o dado em uma dentre duas classes, como por exemplo, se um paciente tem um câncer benigno ou maligno. Para problemas multiclasse, o objetivo é classificar o dado em pelo menos três classes, por exemplo, se uma criança está abaixo, acima ou no peso ideal. E para problema de multirótulo, o objetivo é identificar quais são as classes que estão presentes na instância em questão, por exemplo, dado uma imagem e deseja-se sabe quais objetos estão presentes na imagem, um ou mais objetos podem ser identificados. Neste trabalho, será objeto de estudo, algoritmos de classificação supervisionados para classificações binárias.

### 3.1.1 Avaliação de desempenho para classificação de dados

Na literatura, é possível encontrar métricas que podem ser usadas para avaliar o desempenho de classificadores (Davis e Goadrich, 2006). Entretanto, existem métricas que possuem propriedades importantes para avaliar classificações. Como descrito por Bradley (1997), o *Receiver Operating Characteristic* (ROC) se destaca por permitir sua aplicação em contextos onde há classes desbalanceadas ou poucos dados, além de ser extensivamente usada ao longo dos anos. Além disso, a métrica pode ser visualizada através do gráfico ROC e da curva ROC (Fawcett, 2005).

O gráfico ROC da Figura 1 é composto por dois eixos, taxa de falsos positivos (FP), e taxa de verdadeiros positivos (VP), no eixo x e y, respectivamente. Cada ponto no gráfico representa o resultado de um classificador, onde x = y representa uma baixa pontuação de classificação (neste caso a taxa VP é igual a taxa FP) em contrapartida o ponto (0, 1) representa uma classificação perfeita. A Figura 1 ilustra os possíveis resultados representados no gráfico ROC, onde o ponto D representa uma classificação perfeita, o ponto C representa uma classificação aleatória e o ponto E representa uma classificação muito pior que o aleatório, já o ponto A se mostra mais conservador que o

ponto B. Em muitos problemas reais é preferível casos com taxa de FP mais baixas, portanto o ponto A exemplifica um resultado mais apropriado que B (Fawcett, 2005).

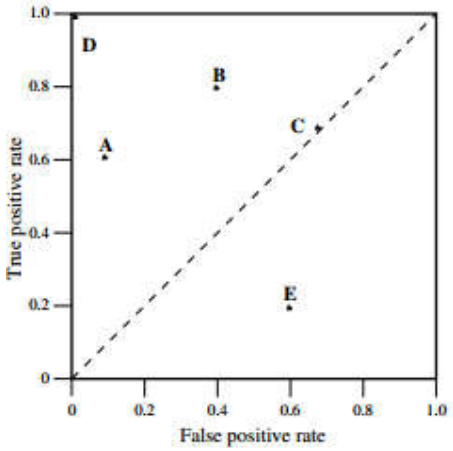


**Figura 1 - Ilustração do gráfico ROC (Fawcett, 2005)**

A curva ROC (Figura 2) é uma alternativa ao gráfico ROC para que se visualize o resultado de classificações binárias onde o resultado apresenta a probabilidade de uma determinada instância pertencer à uma das classes (Fawcett, 2005). Dado que o problema é determinar à qual classe as instâncias pertencem, ao invés da probabilidade em si, se torna importante especificar um limiar para que valores acima do limiar pertençam à uma classe, e abaixo do limiar pertençam à outra classe. Portanto, para cada limiar possível, temos um ponto na curva ROC, e para cada classificador é gerado uma curva, desta forma pode-se observar o desempenho geral dos classificadores. A Figura 2 ilustra o resultado de um classificador e seus respectivos pontos que formam a curva ROC.

| Inst# | Class | Score | Inst# | Class | Score |
|---|---|---|---|---|---|
| 1 | p | .9 | 11 | p | .4 |
| 2 | p | .8 | 12 | n | .39 |
| 3 | n | .7 | 13 | p | .38 |
| 4 | p | .6 | 14 | n | .37 |
| 5 | p | .55 | 15 | n | .36 |
| 6 | p | .54 | 16 | n | .35 |
| 7 | n | .53 | 17 | p | .34 |
| 8 | n | .52 | 18 | n | .33 |
| 9 | p | .51 | 19 | p | .30 |
| 10 | n | .505 | 20 | n | .1 |

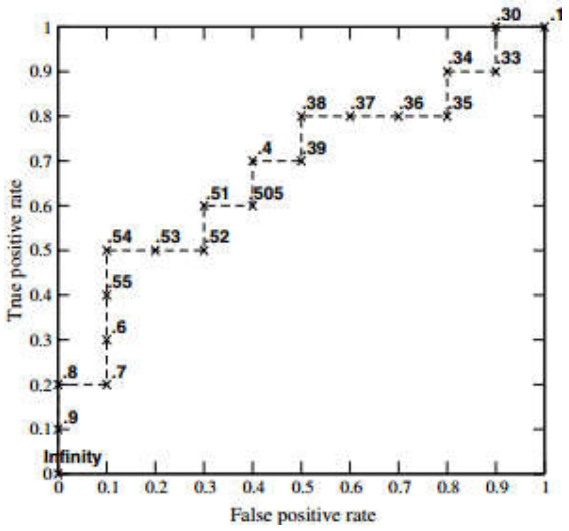


**Figura 2 – Ilustração da curva ROC (Fawcett, 2005)**

Embora a curva ROC represente o desempenho de um classificador de forma gráfica, ainda é uma forma subjetiva para comparar classificadores, pois há muitos pontos a serem analisados. Por isso, é introduzido uma medida conhecida como *Area Under the Curve* (AUC) que calcula a área sob a curva ROC, retornando assim um único valor escalar. Desta forma, valores próximos de 1 correspondem a bons classificadores, e para valores próximos ou abaixo de 0.5 correspondem a classificadores aleatórios. De acordo com Bradley (1997), a métrica AUC é uma boa solução para condensar o desempenho de um classificador em um único escalar. Em seu trabalho, Bradley (1997) apresenta uma série de vantagens que favorecem o uso do AUC, como por exemplo a independência aos limiares, se apresenta um bom indicador de desempenho para

classificação binária e consegue produzir bons resultados mesmo com classes desbalanceadas.

Considerando o experimento realizado nesta monografia, considero que a métrica AUC e suas propriedades se enquadram neste contexto, portanto, para a interpretação de desempenho dos classificadores, este trabalho adotou a métrica AUC como principal métrica de avaliação de desempenho.

## 3.2 Árvores de decisão

As árvores de decisão são uma das abordagens populares na literatura, e muito utilizada em aplicações práticas (Breslow, 1997). De um modo geral, seu funcionamento consiste em tomar um conjunto de dados como entrada e construir uma árvore de nós, onde as folhas determinam a decisão (à qual classe a instância pertence) e os nós representam testes condicionais (acima de um valor, ou abaixo de um valor). Além de apresentar uma boa habilidade em generalizar problemas, uma das vantagens do uso de árvores de decisão, é a possibilidade de interpretar a decisão tomada com base nos nós criados, além dos nós existentes considerarem somente atributos relevantes para a tomada de decisão. Como resumido por Russel e Norvig (2005), "Uma árvore de decisão alcança sua decisão executando uma sequência de testes".

A Figura 3 representa uma árvore de decisão considerando atributos como umidade (*humid*), tempo (*outlook*) e vento (*windy*) para determinar se é possível (P) ou não (N) realizar uma partida de tênis. Para este caso, por exemplo, um dia chuvoso (*rain*) e sem ventos (*windy*) é possível realizar a partida de tênis.

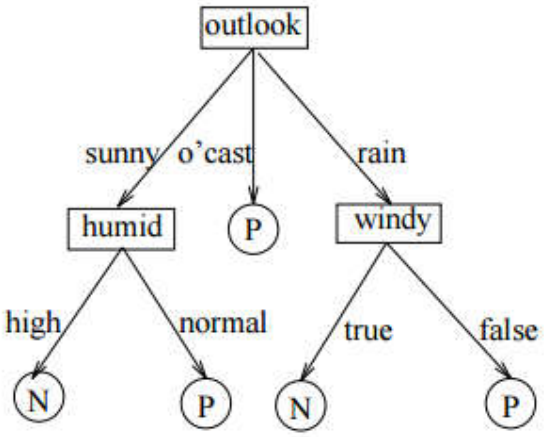


**Figura 3 – Exemplo simplificado de uma árvore de decisão (Quinlan, 1986)**

# 4
# Revisão Bibliográfica

*Este capítulo apresenta alguns trabalhos conhecidos sobre o assunto em questão, abordando métodos de experimentação, imputação de dados e resultados obtidos em experimentos similares.*

## 4.1 Imputação de dados e árvores decisão

Um dos trabalhos específicos sobre imputação de dados em árvores apresentado por Liu *et al.* (1997), o qual introduz o assunto com base em estudos passados mas não apresenta nenhuma solução ao tema. A princípio, é possível imaginar que o processo de imputação em árvores seja a única alternativa para tratar dados faltantes, porém, uma das técnicas exploradas por Feelders (1999), conhecida como *Surrogate Split*. Em seu trabalho, Feelders compara a técnica com a imputação de dados, onde foi possível concluir que a imputação de dados produz resultados mais relevantes. Em um estudo mais direcionado, onde Kalousis e Hilario (2000) examinaram dados faltantes dos tipos MAR e MNAR usando simulações a partir de conjunto de dados reais para analisar as propriedades de sete algoritmos de classificação, dentre eles, dois algoritmos baseados em árvores. Em seu trabalho, concluíram que o algoritmo *Naïve Bayes* foi considerado o mais resistente aos dados faltantes, e os dados faltantes espalhados pelas variáveis impactam mais negativamente quando se comparado com dados faltantes concentrados somente em algumas variáveis. Fujikawa e Ho (2002) apresentaram três algoritmos de imputação baseados em *clusters* utilizando média e moda de complexidade linear, e concluem que os algoritmos apresentados produzem resultados equivalentes aos métodos mais sofisticados analisados em seu trabalho, e portanto podem ser aplicáveis a grandes conjuntos de dados. Batista e Monard (2003) conduziram um experimento comparando quatro algoritmos de imputação e concluíram que o algoritmo de imputação KNN (usando 10 vizinhos) se mostrou superior na maioria dos casos. Em contraste, Kim e Yates (2003) realizaram uma simulação entre sete algoritmos de imputação mais populares, porém não encontraram nenhum método dominante. Zhang

*et al.*, (2005) analisaram a dificuldade de encontrar um valor para um dado faltante, e concluíram que quando o dado faltante exige um alto custo para imputação, é preferível ignorar esses dados, considerando que este fator reduz custos de testes. Saar-Tsechansky e Provost (2007) conduziram um experimento utilizando vários algoritmos de imputação em árvores de decisão e propuseram uma abordagem sensível à custo para o tratamento de dados faltantes onde foi possível obter resultados positivos, porém os dados faltantes estavam presentes somente na fase de teste. E por fim, um experimento mais próximo ao proposto neste trabalho, foi realizado por Ding e Simonoff (2010), onde investigaram métodos de imputação utilizando árvores de decisão em problemas de classificação binária, e concluíram que separar os dados faltantes em uma classe específica produz melhores resultados em relação aos demais imputadores.

Muitos trabalhos assumem o tipo de dados faltantes conhecido como MCAR, pois grande parte dos algoritmos de imputação assumem este tipo para operar. Neste trabalho, também será assumido o tipo MCAR para os dados faltantes que serão artificialmente gerados a partir de conjunto de dados completos. Esta abordagem de inserção artificial dos dados é similar à adotada por Kalousis e Hilario (2000). Em consideração à ambientes práticos, onde o conjunto de treinamento e teste podem conter dados faltantes, este trabalho será considerado da mesma forma, contrapondo o experimento conduzido por Saar-Tsechansky e Provost (2007). Como sugerido por Kalousis e Hilario (2000), neste trabalho os dados faltantes gerados artificialmente serão espalhados aleatoriamente por todas as variáveis do conjunto de dados, visando compreender o caso onde produz o maior impacto nos resultados. Dessa maneira, considero ser possível obter resultados mais realistas e robustos, dado que assim é possível considerar o caso extremo de dados faltantes. Diferentemente dos trabalhos mencionados, este trabalho busca levar em consideração estudos onde não há intervenção de algoritmos de imputação, onde o próprio algoritmo de classificação é responsável por tratar internamente os dados faltantes, para que assim, seja possível comparar com os resultados gerados pelos algoritmos de imputação.

# 5
# Metodologia

*Este capítulo apresenta a metodologia utilizada na monografia. Na Seção 5.1 apresenta uma visão geral do experimento descrevendo brevemente os processos principais. Nas seções seguintes, é detalhado cada uma das fases do processo de experimentação.*

## 5.1 Processo geral de experimentação

Para a realização do experimento, alguns passos foram estabelecidos para sistematizar a execução. Como ilustrado na Figura 4, os passos podem ser descritos em três fases, inserção de dados faltantes, imputação de dados e classificação de dados. O resultado de cada fase é passado como entrada para a fase seguinte. Cada fase tem um objetivo especifico, na inserção de dados faltantes a tarefa é produzir amostras inserindo valores nulos de forma aleatória nos colunas em proporções graduais em cada conjunto de dados, na fase seguinte a principal função é preencher os valores nulos com diferentes algoritmos de imputação. Por fim a última fase do experimento busca classificar todos os conjuntos de dados para que seja possível traçar comparativos entre algoritmos de imputação e classificação, além de analisar a perda de precisão entre cada nível estipulado.

As entradas escolhidas para serem estudadas neste trabalho foram baseadas em três critérios: todos os conjuntos de dados devem ser conhecidos na literatura, devem conter somente dados numéricos e não deve haver nenhum dado faltante. A popularidade dos conjuntos de dados é definida pela quantidade de *downloads* no repositório e pela citação em publicações científicas, e os demais critérios foram estabelecidos para definir escopo ao experimento. A origem dos conjuntos de dados apresentados neste trabalho foram extraídos do repositório *UCI Machine Learning Repository*. Além disso, é importante notar que os dados foram classificados sem nenhuma alteração ou pré-processamento.

Os cinco conjuntos de dados escolhidos são de informações retiradas de diferentes contextos, mas todos respeitando os critérios estabelecidos para este trabalho. O primeiro conjunto de dados é conhecido como *Diabetic Retinopathy Debrecen Data Set* (Antal e Hadju, 2014), o qual contém 19 características extraídas de 1157 imagens do conjunto *Messidor* (Klein *et al.*, 2016) para classificar quando uma imagem contém sinais de retinopatia diabética ou não. O segundo conjunto de dados é conhecido como *Blood Transfusion Service Center Data Set* (Yeh, Yang e Ting, 2009), o qual contém 5 características referentes à 748 doadores de sangue na qual busca classificar se o doador doou sangue em Março de 2007 ou não. O terceiro conjunto de dados é conhecido como *Breast Cancer Wisconsin (Original) Data Set* (Mangasarian, Street e Wolberg, 1995), contendo 9 características de amostras mamárias de 699 pacientes, no qual o objetivo é classificar se o paciente tem ou não câncer de mama. O quarto conjunto de dados é conhecido como *Ionosphere Data Set* (Sigillito *et al.*, 1989), contendo 34 características sobre elétrons livres na ionosfera e 351 registros, com o objetivo de identificar evidência (ou não) de algum tipo de estrutura na ionosfera. E por último, o *SPECTF Heart Data Set* (Kurgan *et al.*, 2001) com 44 características de 267 pacientes, descrevendo o diagnóstico cardíaco através de imagens de tomografia computadorizada por emissão de fóton único, onde se busca classificar a condição do paciente em "normal" ou "anormal".

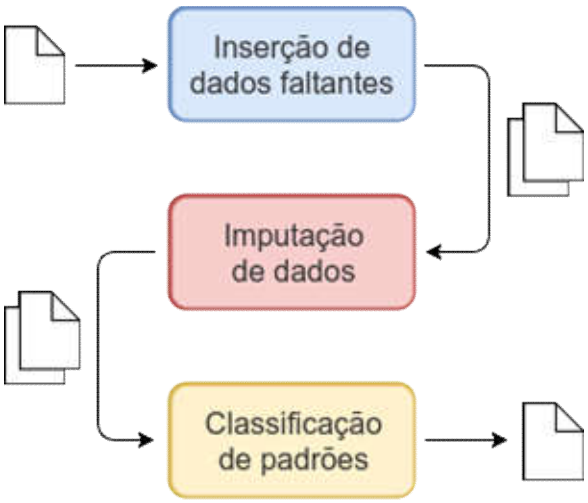


**Figura 4 – Ilustração do processo geral de experimentação.**

## 5.2 Inserção de dados faltantes

A primeira fase é conhecida como inserção de dados faltantes, a qual consiste em obter o conjunto de dados original (sem nenhum dado faltante) e inserir valores nulos aleatoriamente em diferentes níveis percentuais, partindo de 10% até 90%, com um incremento de 10%, como ilustrado na Figura 5. Ao inserir os primeiros valores aleatórios (referente à 10%) um novo conjunto de dados com dados faltantes é criado. No próximo nível (referente à 20%), o conjunto de dados gerado no nível anterior (referente à 10%) é tomado como entrada, ou seja, os valores inseridos como nulo no nível anterior são aproveitados neste nível, e novos 10% de valores nulos são inseridos para totalizar os 20%. Para os níveis seguintes, a mesma regra é respeitada até que se complete os 90%. Uma vez completado todos os níveis, temos então a primeira amostragem, onde cada amostragem contém 9 conjuntos de dados faltantes referentes a cada nível de percentual.

O uso dessa abordagem, onde em cada amostragem é inserido gradualmente 10% de valores nulos em posições aleatórias para cada nível, permite que seja assumido o tipo MCAR para os conjuntos de dados gerados. Esse processo de amostragem é realizado 10 vezes, de modo a produzir casos suficientes para obter resultados mais consistentes e favorecer uma análise com maior representatividade. A metodologia adotada para a inserção de dados faltantes é baseada no trabalho desenvolvido por Kalousis e Hilario (2000).

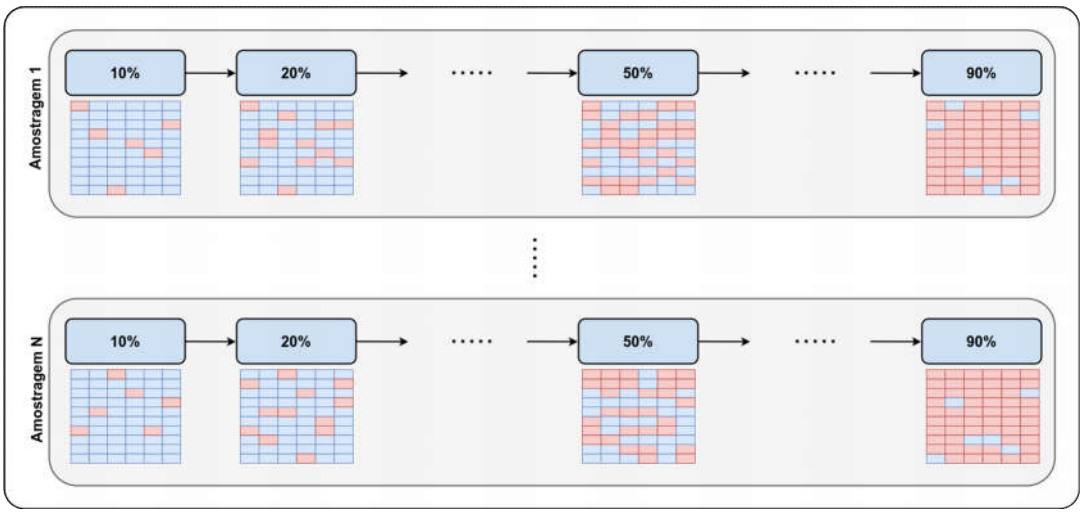


**Figura 5 – Ilustração do processo de inserção de dados faltantes.**

## 5.3 Imputação de dados

A segunda fase é onde ocorre a imputação dos dados, e como entrada de dados se obtém todos os conjuntos de dados de cada amostragem gerados na fase anterior (Inserção), e então busca "preencher" os valores nulos de cada conjunto em cada amostragem. Para a seleção dos algoritmos de imputação usados neste trabalho, foi determinado como critérios a popularidade na literatura e disponibilidade em pacotes de softwares, e seu mecanismo seja bem documentado.

Para cada conjunto de dadas faltante, cinco algoritmos de imputação de forma independente imputam os dados faltantes de modo a produzir conjuntos de dados completos. Os algoritmos usados para esta fase são: *Expectation Maximization* (EM) (Schager, 1997), *Multiple Imputation by Chained Equation* (MICE) (Buuren e Groothuis-Oudshoorn, 2011), *K-Nearest Neighbors* (KNN) (Rubinsteyn e Feldman, 2017), media (MEAN), mediana (MEDIAN) e NONE onde não há imputação de dados, mas o próprio algoritmo de classificação tratará os dados faltantes.

As implementações utilizadas para MICE e KNN são dos pacotes *fancyimpute* (Rubinsteyn e Feldman, 2017) na linguagem Python, MEAN e MEDIAN foram usadas da biblioteca *scikit-learn* (Buitinck *et al.*, 2013) na linguagem Python, e EM do pacote *Weka* (Hall *et al.*, 2013) na linguagem Java. Para todos algoritmos, foi utilizado os parâmetros padrões de cada implementação (quando aplicado), com exceção do algoritmo KNN, onde foi definido o parâmetro de 5 vizinhos. Devido a heterogeneidade das linguagens utilizadas neste experimento, foi adotado a biblioteca *python-weka-wrapper* (Reutemann, 2017) para que fosse possível utilizar os pacotes em linguagem Java através da linguagem Python, assim todo o processo de imputação e experimentação foi desenvolvido na linguagem Python.

## 5.4 Classificação de dados

A terceira e última fase, é a classificação de dados, onde consiste em tomar os conjuntos de dados gerados na fase anterior, e classifica-los com os algoritmos de classificação

determinados para este experimento. Os algoritmos utilizados nesta fase são o C.45 (Quinlan, 1993) e o *Random Forest* (Breiman, 2001), com as implementações existentes no software *Weka* (Hall *et al*., 2009), utilizando os parâmetros padrão de cada algoritmo. É importante ressaltar que o algoritmo C.45 no *software Weka* é implementado no algoritmo J48. Assim como nas fases anteriores, foi utilizado o pacote *python-weka-wrapper* (Reutemann, 2017) para que fosse possível utilizar os recursos do *software Weka* na linguagem Python.

O processo de classificação consiste em 3 execuções independentes, e para cada execução a forma de validação é feita através da validação cruzada conhecida como *k-fold* (Geisser, 1975), a qual divide o conjunto de dados em 5 partes e escolhe aleatoriamente quatro partes para o treinamento e uma para o teste, e alterna este processo por 5 vezes. Ao fim deste processo, é emitido um relatório detalhado de cada parte da validação e das execuções independentes.

# 6
# Resultados

*Este capítulo apresenta os resultados obtidos no experimento, e descreve os resultados através de três análises específicas. Na Seção 6.1 é discutido os resultados quanto ao desempenho geral. Na Seção 6.2 é analisado o comportamento dos classificadores quando não há utilização de imputadores. Na Seção 6.3 é analisado o decaimento da pontuação AUC após 50% de dados imputados.*

As descrições das análises serão apresentadas em três perspectivas, onde a primeira apresentará uma visão geral do desempenho com seus devidos comentários, em seguida uma comparação entre os algoritmos de imputação será através de gráficos de linha, destacando os casos de interesse, e finalmente uma análise do decaimento da pontuação AUC para os diversos casos deste experimento.

## 6.1 Análise de desempenho geral

Nas Figuras de Figura 6 a Figura 15 são apresentados gráficos *boxplot* contendo as informações do desempenho geral de cada conjunto de dados e algoritmo de classificação. Na organização das figuras, são apresentados dois gráficos por conjunto de dados, contendo os resultados da pontuação média AUC de cada classificador. E por fim, as discussões e interpretações de cada figura serão apresentadas pontualmente.

As Figuras Figura 6 e Figura 7, são casos que considero equivalentes, pois apresentam resultados similares e o imputador MEDIAN apresenta os piores resultados para os dois classificadores. Para estes casos, não vejo uma situação em que um imputador seja superior aos demais, apenas que o classificador J48 apresenta uma pontuação AUC média inferior ao *Random Forest*.

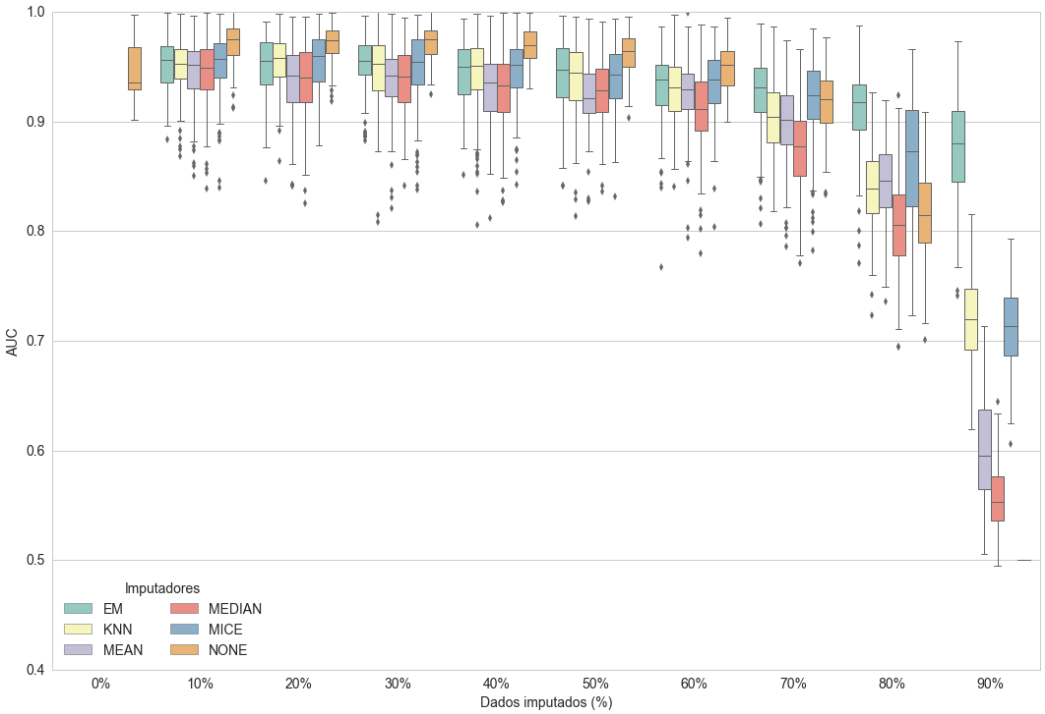


**Figura 6 - Resultados da classificação em "breast-cancer " usando J48**

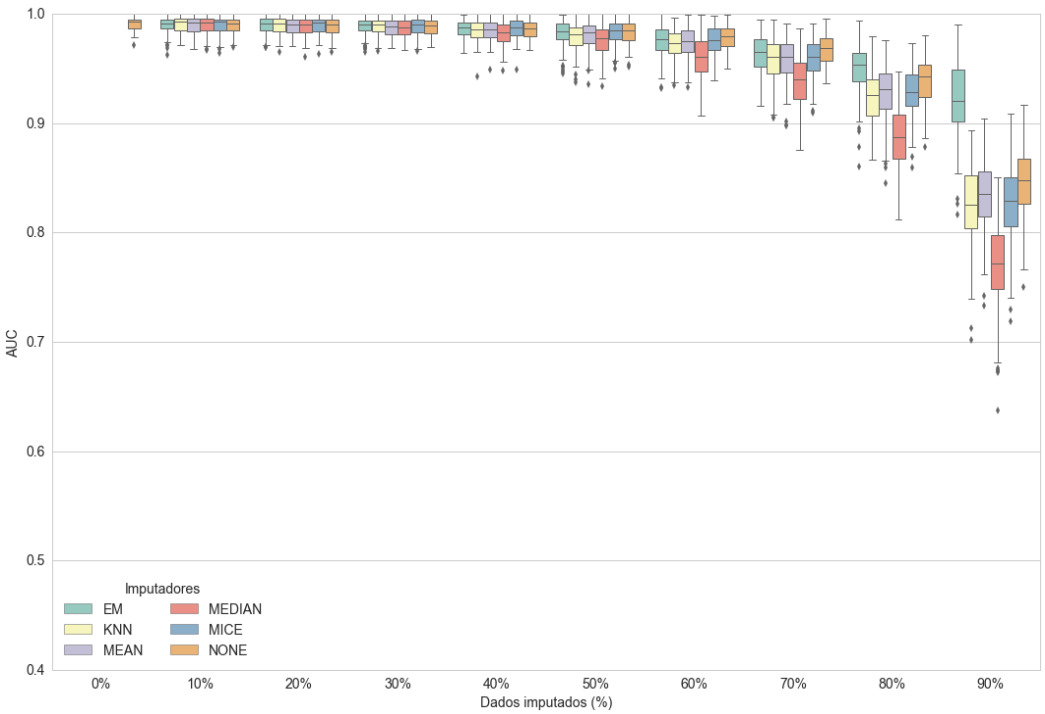


**Figura 7 - Resultados da classificação em "breast-cancer " usando Random Forest**

Para o conjunto de dados "ionosphere" (Figura 8 e Figura 9), observei que ao utilizar o próprio classificador J48 para tratar os dados faltantes (imputador NONE na Figura 8) se obtém uma pontuação ligeiramente superior ao imputador MICE (Figura 8). Em contraste, observo que ao utilizar o classificador *Random Forest* para tratar os dados faltantes (Imputador NONE na Figura 9) se obtém uma pontuação ligeiramente inferior ao imputador MICE (Figura 9). Considero então, como um onde o mesmo imputador apresenta resultados diferentes em um mesmo conjunto de dados.

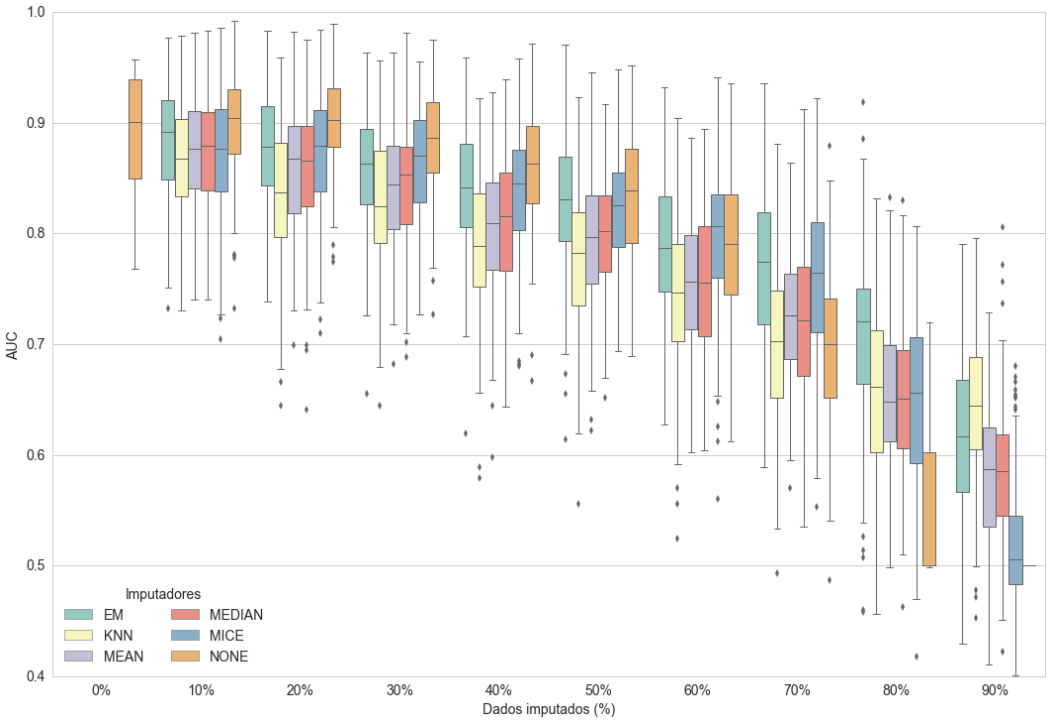


**Figura 8 - Resultados da classificação em "ionosphere" usando J48**

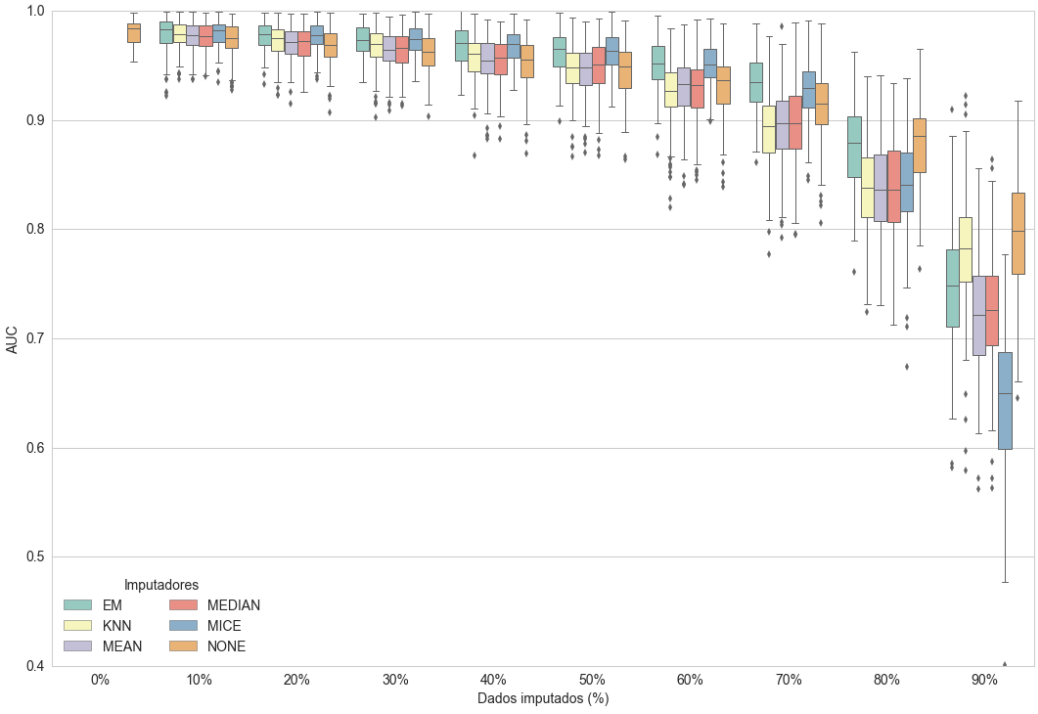


**Figura 9 - Resultados da classificação em "ionosphere" usando Random Forest**

Quanto a determinação de imputadores dominantes, vejo em casos específicos, a existência imputadores que apresentam resultados mais significativos que outros, como no caso do conjunto de dados “messidor” (Figura 10 e Figura 11) onde os imputadores MICE e EM, exibem resultados mais significativos em relação aos demais imputadores.

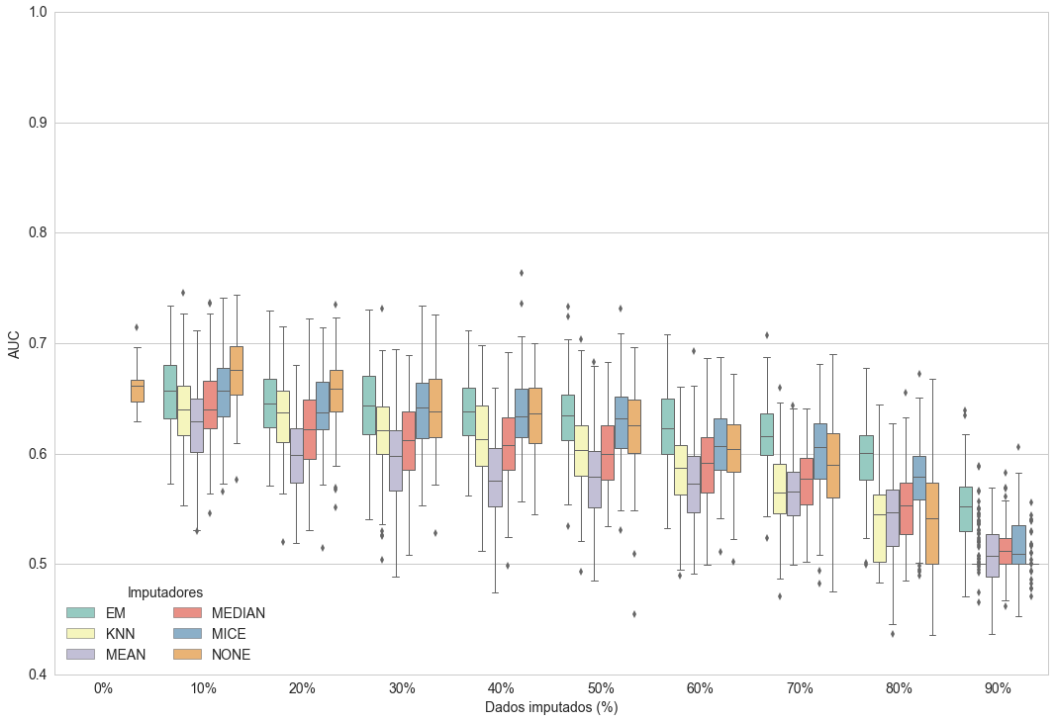


**Figura 10 - Resultados da classificação em "messidor" usando J48**

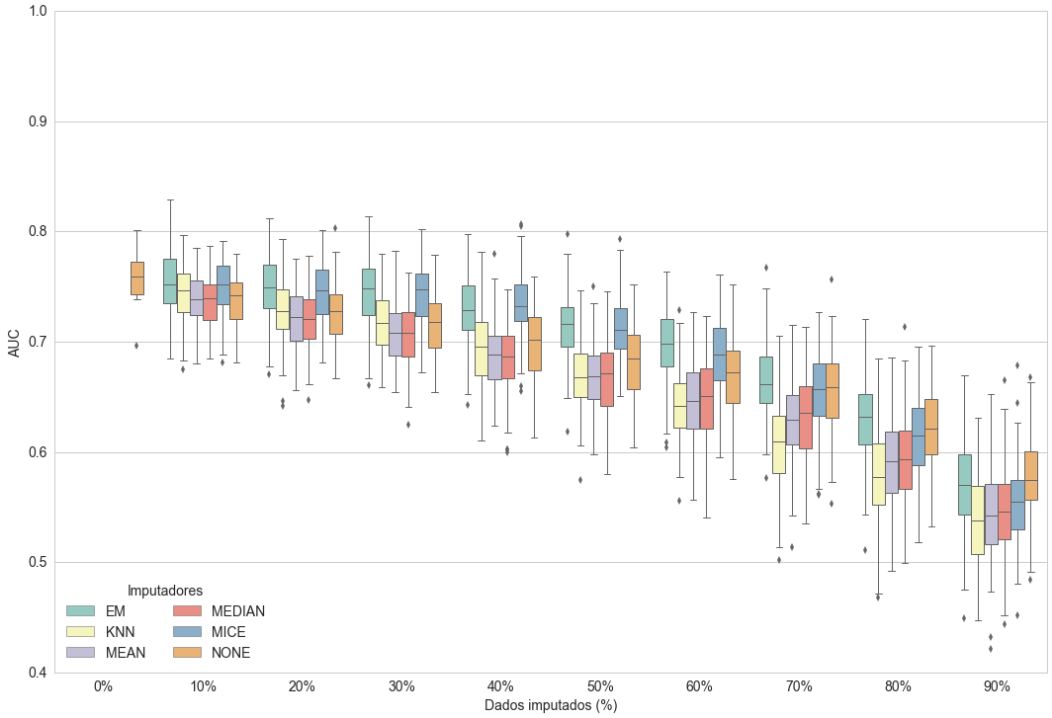


**Figura 11 - Resultados da classificação em "messidor" usando Random Forest**

Em contraste aos casos anteriores, também notei que há casos onde não é possível distinguir qual imputador se sobressai, como por exemplo no conjunto de dados “spectf” (Figura 12 e Figura 13), na qual todos os imputadores apresentam resultados similares para qualquer percentual de imputação e classificador utilizado.

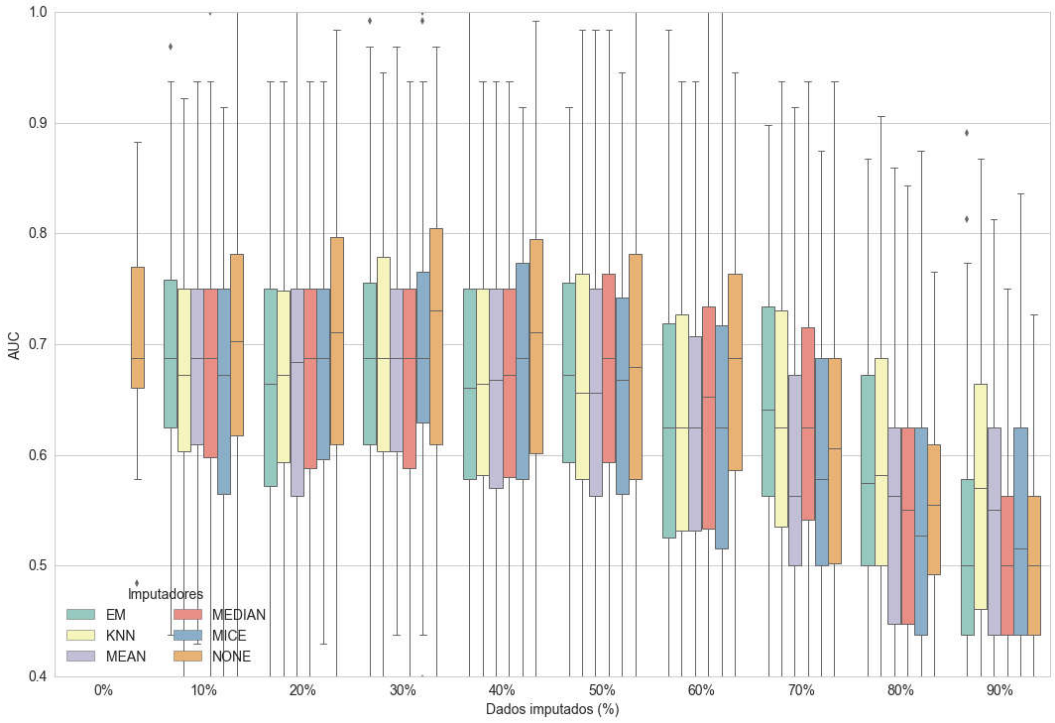


**Figura 12 - Resultados da classificação em "spectf" usando J48**

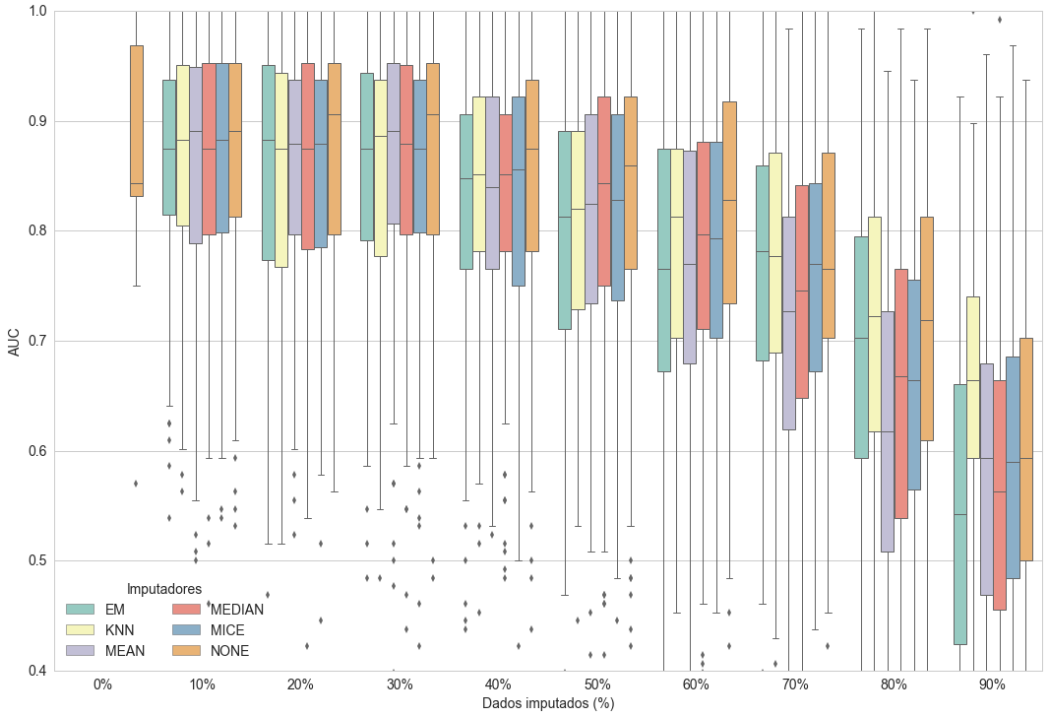


**Figura 13 - Resultados da classificação em "spectf" usando Random Forest**

Nas Figuras Figura 14 e Figura 15, são exemplos de situações onde não consigo identificar um imputador que seja superior aos demais para a maioria dos casos. Além disso, noto que para a Figura 14 após 30% de dados imputados, o classificador e os imputadores não são capazes de distinguir as classes com grande precisão. Entendo que este comportamento, pode ser atribuído à uma perda de dados considerados relevantes para a determinação das classes.

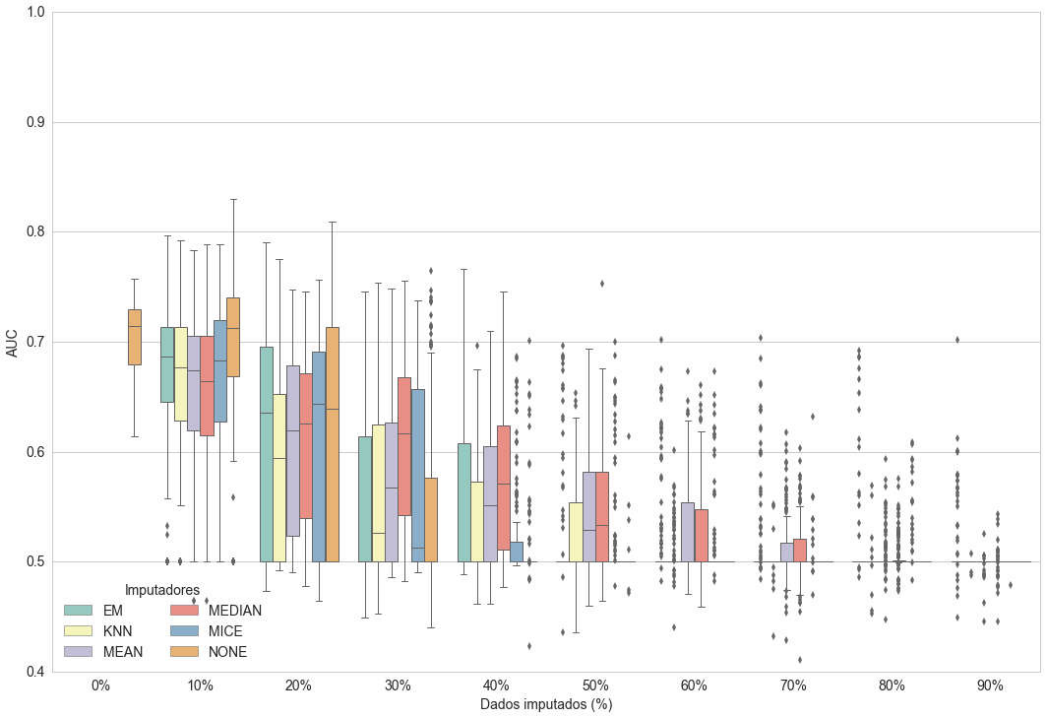


**Figura 14 - Resultados da classificação em "transfusion" usando J48**

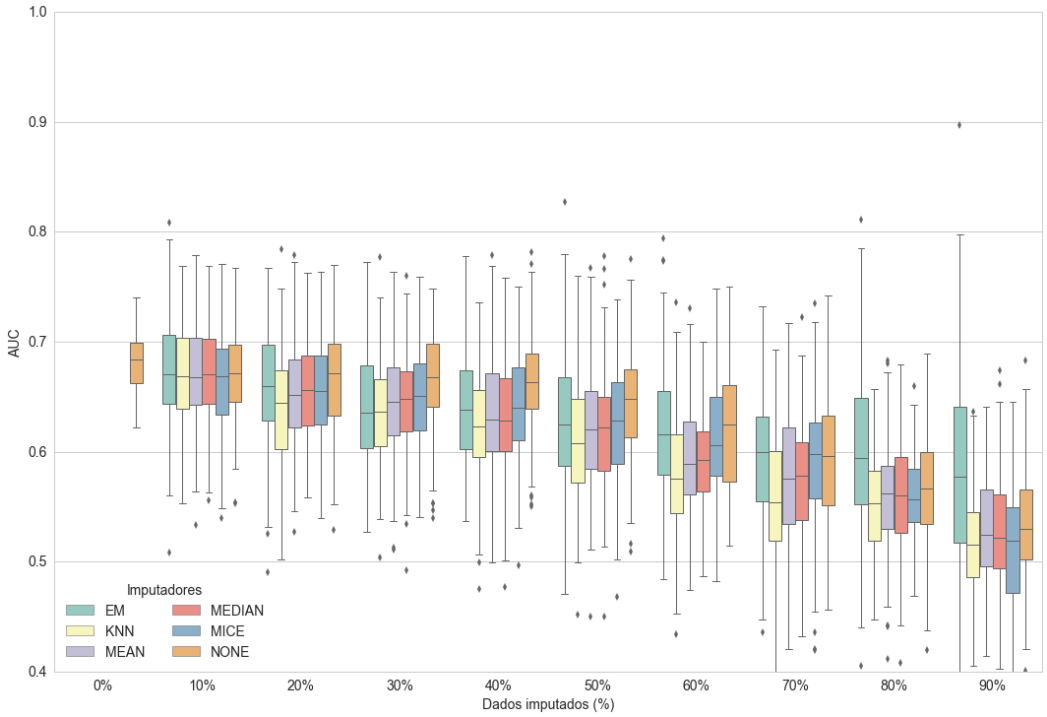


**Figura 15 - Resultados da classificação em "transfusion" usando Random Forest**

Nas figuras apresentadas anteriormente, de um modo mais geral identifiquei que o classificador J48 produz pontuações AUC baixas, e se demonstra mais sensível aos dados faltantes em relação ao classificador *Random Forest*. Em ambos dos classificadores, é possível observar uma perda gradual na pontuação AUC na medida que o percentual de dados imputados aumenta.

Embora tenha observado as diferenças comentadas para cada gráfico, considero impróprio afirmar a existência de imputadores que contribuam para a recuperação dos dados faltantes, dado que os experimentos neste trabalho foram conduzidos com um conjunto limitado de ferramentas. Entretanto, com base nos resultados apresentados neste experimento, noto que há comportamentos divergentes dependendo do conjunto de dados, imputador e classificador utilizado, o que sugere uma dependência entre os objetos de estudo onde os quais podem produzir resultados variados de acordo com o contexto. Portanto não considero que há algum imputador predominante nos contextos apresentados neste trabalho. Esta conclusão, embora os imputadores analisados sejam diferentes, confirma os resultados obtidos por Kim e Yates (2003), nos quais também não identificaram a existência de um imputador que fosse superior em todos os casos estudados.

Um fator que considero relevante em ser ressaltado, é a perda na pontuação AUC mais acentuada após 50% de dados imputados para todos os casos. Este fator é analisado e discutido com mais detalhes na Seção 6.3 utilizando gráficos mais adequados para a interpretação desta observação.

## 6.2 Usando classificadores para tratar dados faltantes

Como apresentado inicialmente neste trabalho, existem classificadores capazes tratar dados faltantes, sem o auxílio de imputadores para prever os dados nulos do conjunto de dados. Esse tipo de abordagem pode ser útil para poupar tempo durante o processo de modelagem em diversos problemas. Nesta seção, será apresentado um comparativo dos resultados da média da pontuação AUC em gráficos de linha de cada conjunto de dados, de modo a contrastar os resultados dos demais imputadores em comparação ao

próprio algoritmo de classificação. Dessa forma, pode ser possível obter sugestões de contextos onde utilizar o classificador para tratar dados faltantes seja mais adequado.

Para interpretação dos gráficos, o imputador NONE é referente ao gráfico do classificador. Na Figura 16 por exemplo, o gráfico a esquerda ilustra os resultados do classificador J48 e a direita ilustra os resultados do classificador *Random Forest*, a curva em destaque (de cor vermelha) representa o imputador NONE, onde em cada gráfico ilustra o próprio classificador em questão tratando os dados faltantes. A mesma interpretação é seguida nas demais figuras.

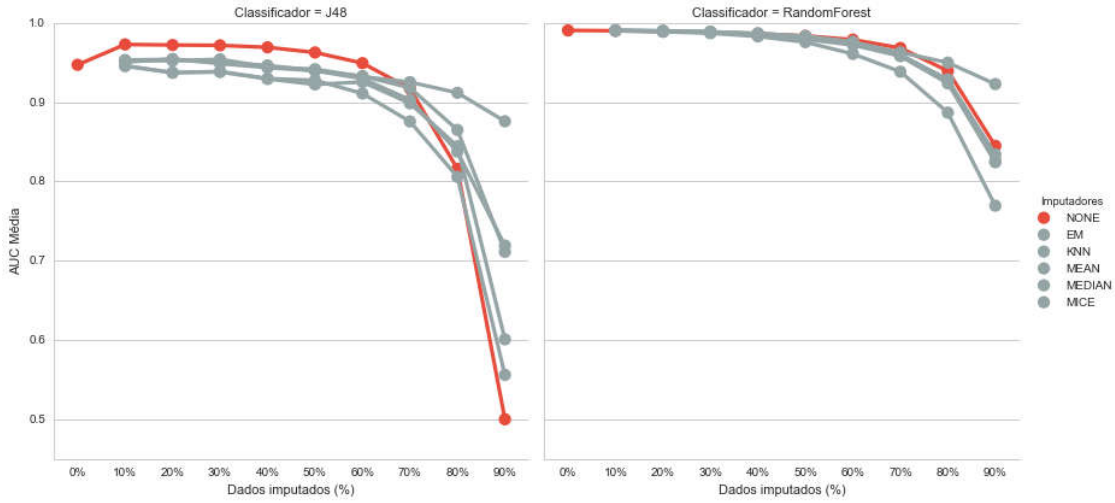


**Figura 16 - Resultado da pontuação AUC média do conjunto de dados "breast-cancer"**

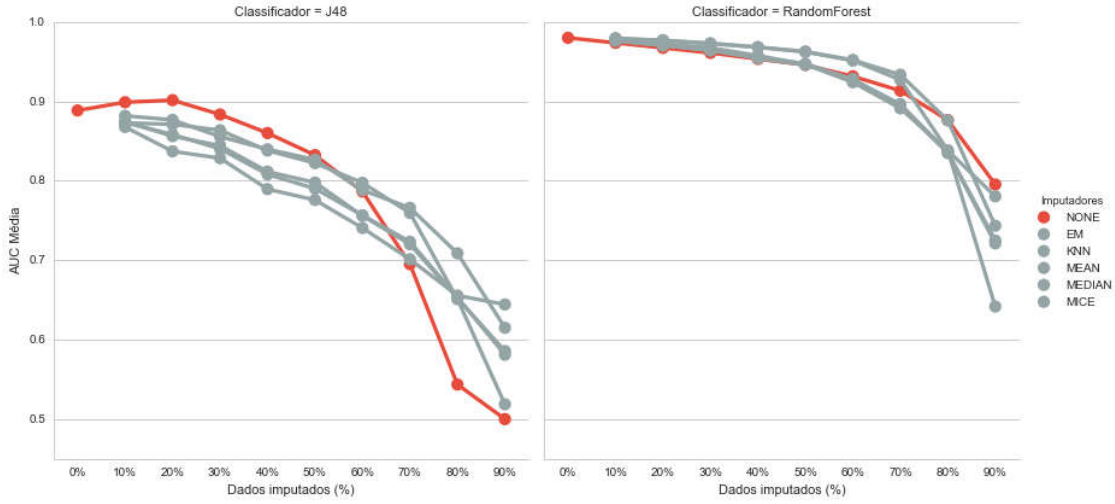


**Figura 17 - Resultado da pontuação AUC média do conjunto de dados "ionosphere"**

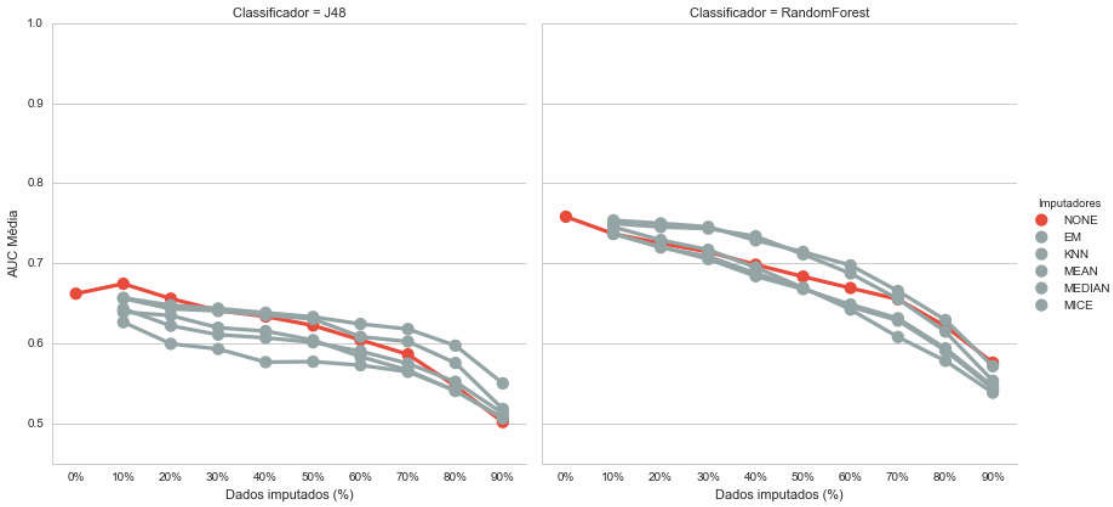


**Figura 18 - Resultado da pontuação AUC média do conjunto de dados "messidor"**

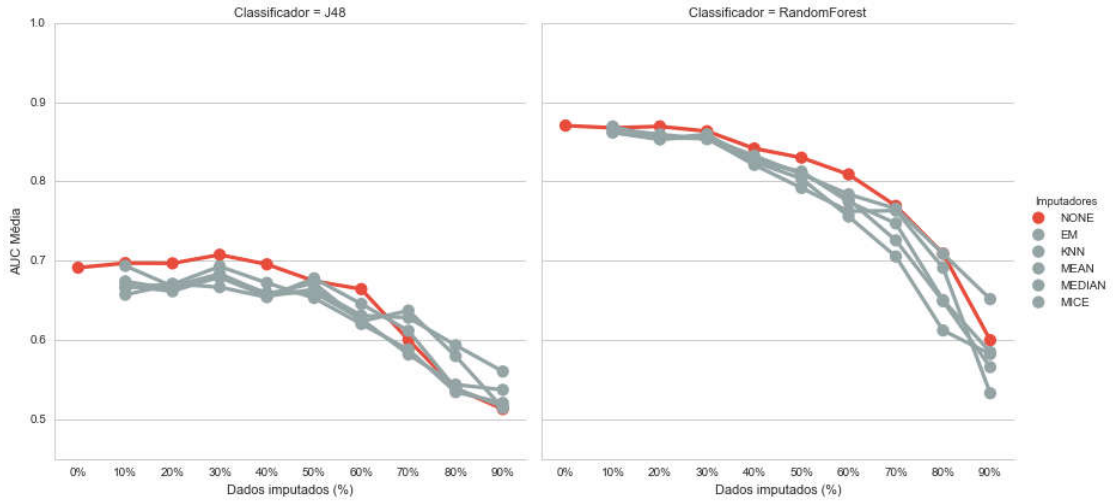


**Figura 19 - Resultado da pontuação AUC média do conjunto de dados "spectf"**

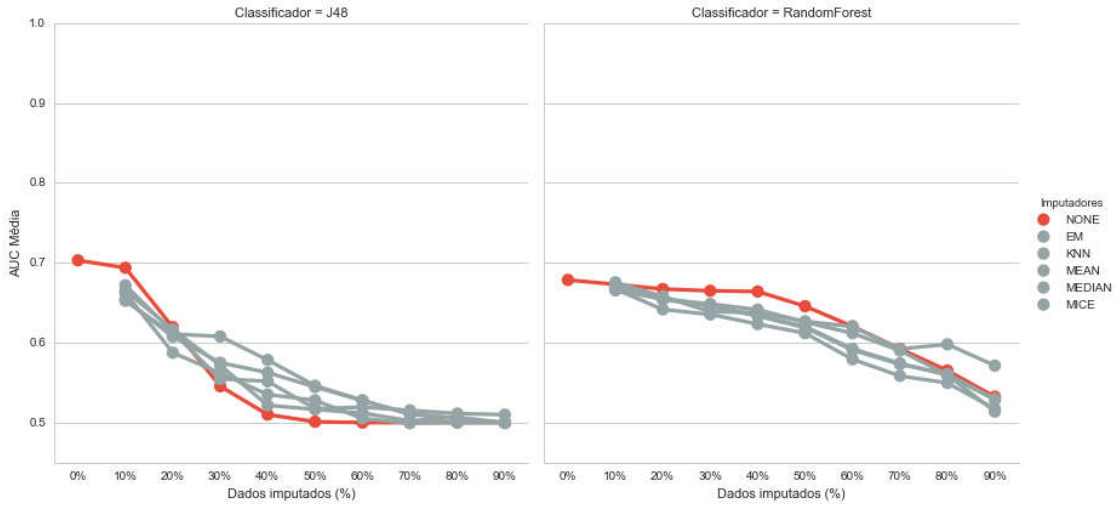


**Figura 20 - Resultado da pontuação AUC média do conjunto de dados "transfusion"**

Considerando somente os gráficos referentes ao J48, observo que em até 50% de dados imputados o imputador NONE consegue lidar com os dados faltantes de maneira similar aos demais imputadores, com exceção do conjunto de dados “transfusion”, onde os imputadores conseguem obter resultados superiores. Além disso, vejo que após o limiar de 50%, a utilização de imputadores pode contribuir significativamente ao invés de utilizar somente o classificador para tratar os dados faltantes.

De forma similar, considerando somente os gráficos referentes ao *Random Forest*, observo que ao utilizar o próprio classificador para tratar dados faltantes independente do percentual de dados faltantes, produz resultados similares quando usado imputadores. Dessa forma, vejo como uma interpretação divergente ao que foi levantado em relação ao classificador J48, onde o uso de imputadores é condicional ao percentual de dados faltantes.

Em suma, observo que o classificador J48 devido a sua simplicidade se mostra mais sensível à dados faltantes em comparação ao classificador *Random Forest*, e portanto me parece prudente a utilização de imputadores para minimizar a perda de resultados à partir de 50% de dados faltantes. Em contraste, o classificador *Random Forest* consegue produzir resultados equivalentes quando utilizado imputadores para o tratamento de dados faltantes, o que pode ser uma boa opção para modelagem de problemas dado que o processo de imputação pode ser eliminado.

## 6.3 Análise de decaimento da pontuação AUC

Como apontado na Seção 6.1, observei que há uma queda significativa na pontuação AUC após 50% de dados imputados. Portanto, nesta seção o objetivo é buscar compreender melhor este comportamento.

Na Figura 21, temos uma visão geral de todos os conjuntos de dados presentes neste experimento visando destacar o resultado da pontuação AUC de cada classificador independente do imputador utilizado com uma margem de confiança de 95% no intervalos de 0% a 50% e 50% a 90% de dados imputados.

Em uma análise mais genérica, observo que na maioria dos casos a queda da pontuação AUC após 50% de dados imputados é significativa para qualquer classificador e conjunto de dados utilizado neste trabalho, exceto no conjunto de dados "transfusion" onde o classificador J48 tem uma queda mais significativa antes de 50%, ao contrário dos demais casos. Outro fator, é o intervalo de confiança das curvas, onde na maioria dos casos, os resultados apontam para próximo à média da curva, indicando que este comportamento é predominante na maioria das amostragens. Embora o conjunto de dados "breast-cancer" apresente uma margem maior no intervalo de confiança, o comportamento de queda ainda é predominante assim como nos demais gráficos.

Com esta observação, entendo que existe uma queda na pontuação AUC após 50% de dados imputados, e portanto este fator deve ser considerando durante a manipulação de dados faltantes independente da abordagem utilizada. Uma justificativa que considero plausível para este comportamento, é que após este limiar de 50%, existe um percentual maior de dados artificiais que dados reais presente no conjunto de dados. Dessa forma, os classificadores perdem informações reais que seriam relevantes para uma imputação e até mesmo uma classificação mais precisa.

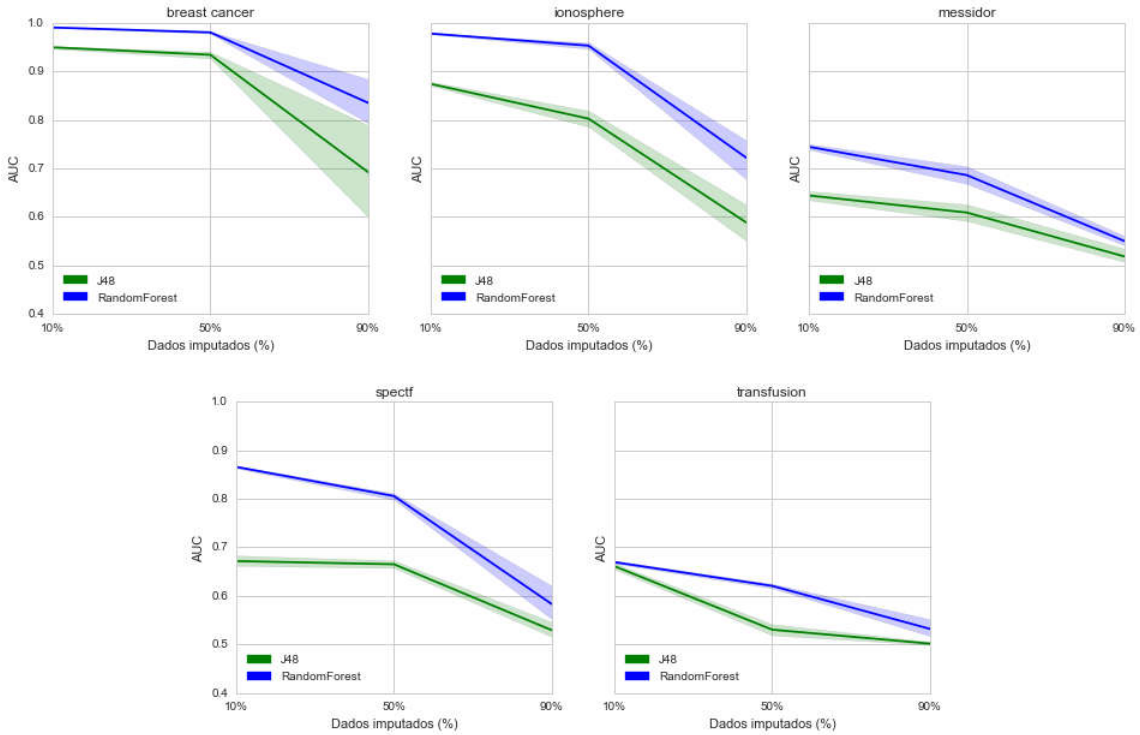


**Figura 21 - Gráfico ilustrando a pontuação AUC em 0%, 50% e 90% de dados imputado**

# 7
# Conclusões e trabalhos futuros

*Este capítulo apresenta as conclusões do trabalho. A Seção 7.1 apresenta as conclusões obtidas pelo experimento deste trabalho. A Seção 7.2 apresenta algumas sugestões de trabalhos futuros.*

## 7.1 Conclusões

Considerando os resultados obtidos através dos experimentos realizados, e o tipo de dados faltantes assumido (MCAR) neste trabalho, as conclusões que obtive podem ser pontuadas em dois aspectos. O primeiro, para classificadores mais simples e sensíveis à dados faltantes como J48, a utilização de imputadores em casos com mais de 50% de dados faltantes, pode minimizar a perda de qualidade na classificação. E segundo, para classificadores mais robustos e tolerantes à dados faltantes como *Random Forest*, a utilização de imputadores pode ser desnecessária dado que o próprio classificador consegue obter bons resultados sem o auxílio de imputadores. Portanto, devido aos diferentes resultados apresentados neste trabalho e discutidos na Seção 6.1, não é possível determinar um método de imputação que seja superior aos demais, esta conclusão é compatível com os resultados de Kalousis e Hilario (2000). Os demais autores revisados neste trabalho, em alguns casos conseguem encontrar imputadores com desempenho superior à outros, porém não tratam do assunto de modo genérico como é proposto neste trabalho.

## 7.2 Trabalhos futuros

- Analisar impacto em classificadores lineares
- Realizar experimento considerando variáveis categóricas
- Explorar algoritmos de imputação alternativos
- Realizar experimento considerando um número maior de conjuntos de dados e apresentar conclusões estatísticas
- Analisar imputadores assumindo outros tipos de dados faltantes

# 8
# Referências Bibliográficas


Aggarwal, Charu C., *Data classification: algorithms and applications*. CRC Press, 2014.

Antal, B. e Hajdu, A. *An ensemble-based system for automatic screening of diabetic retinopathy*. Knowledge-Based Systems, 2014, Vol. 60, pp.20-27

Allison, P.D. *Missing data: Quantitative applications in the social sciences*. British Journal of Mathematical and Statistical Psychology, 2002, Vol. 55(1), pp.193-196

Batista, G.E. e Monard, M.C. An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence, 2003, Vol. 17(5-6), pp.519-533

Bradley, A.P. *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. Pattern recognition, 1997, Vol. 30(7), pp.1145-1159

Breiman, L. *Random forests*. Machine learning, 2001, Vol. 45(1), pp.5-32

Breslow, L.A. e Aha, D.W. *Simplifying decision trees: A survey*. The Knowledge Engineering Review, 1997, Vol. 12(1), pp.1-40

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J. e Layton, R. *API design for machine learning software: experiences from the scikit-learn project*, arXiv preprint arXiv:1309.0238, 2013

Buuren, S. e Groothuis-Oudshoorn, K. *MICE: Multivariate imputation by chained equations in R*. Journal of statistical software, 2011, Vol. 45(3)

Davis, J., e Goadrich, M. *The relationship between Precision-Recall and ROC curves*. Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 233-240

Ding, Y. e Simonoff, J.S. *An investigation of missing data methods for classification trees applied to binary response data*. Journal of Machine Learning Research, 2010, Vol. 11(Jan), pp.131-170

Fawcett, T. *An introduction to ROC analysis*. Pattern recognition letters, 2006, Vol. 27(8), pp.861-874

Feelders, A. *Handling missing data in trees: surrogate splits or statistical imputation?* European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 1999, pp. 329-334

Fujikawa, Y., Ho e T. May, *Cluster-based algorithms for dealing with missing values*. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2002, pp. 549-554

Geisser, S. *The predictive sample reuse method with applications*. Journal of the American statistical Association, 1975, Vol. 70(350), 320-328.

Gelman, A. e Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, H. I. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations. 2009

Heitjan, D.F. e Rubin, D.B. *Ignorability and coarse data*. The annals of statistics, 1991, pp.2244-2253

Kalousis, A. e Hilario, M. *Supervised knowledge discovery from incomplete data*. WIT Transactions on Information and Communication Technologies, 2000, Vol. 25

Klein, J.C., Menard, M., Cazuguel, G., Fernandez-Maloigne, C., Shaefer, G., Gain, P., Cochener, B., Massin, P., Lay, B. e Charton, B. *Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR)*. Ecole des Mines de Paris, 2016

Kim, H. e Yates, S. *Missing value algorithms in decision trees*. Statistical Data Mining and Knowledge Discovery. Chapman and Hall/CRC, 2003

Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M. e Goodenday, L.S. *Knowledge discovery approach to automated cardiac SPECT diagnosis*. Artificial intelligence in medicine, 2001, Vol. 23(2), pp.149-169

Little, R.J. *A test of missing completely at random for multivariate data with missing values*. Journal of the American Statistical Association, 1988, Vol. 83(404), pp.1198-1202

Liu, W.Z., White, A.P., Thompson, S.G. e Bramer, M.A. *Techniques for dealing with missing values in classification*. Advances in Intelligent Data Analysis. Reasoning about Data: Second International Symposium, Springer Berlin/Heidelberg, 1997, p. 527

Mangasarian, O.L., Street, W.N. e Wolberg, W.H. *Breast cancer diagnosis and prognosis via linear programming*. Operations Research, 1995, Vol. 43(4), pp.570-577

Pellegrin, P. *Aristotle's classification of animals: biology and the conceptual unity of the Aristotelian corpus*. University of California Press, 1986

Quinlan, J.R. *Induction of decision trees*. Machine learning, 1986, Vol. 1(1), pp.81-106.

Quinlan, J.R. *C4.5: Programs for Machine Learning*. Morgan Kaugmann Publishers, 1993

Reutemann, P. *"python-weka-wrapper"*. Python Package Index, 04 Jan. 2017, Acesso 02 Abr. 2017. <https://pypi.python.org/pypi/python-weka-wrapper>

Rubinsteyn, A. e Feldman, S. *"fancyimpute"*. Python Package Index, 24 Fev. 2017, Acesso 03 Abr. 2017. <https://pypi.python.org/pypi/fancyimpute>

Russell, S. e Norvig, P. *AI a modern approach*. Learning, 2005, Vol. 2(3), p.4

Saar-Tsechansky, M. e Provost, F. *Handling missing values when applying classification models*. Journal of machine learning research, 2007, Vol. 8(Jul), pp.1623-1657

Sigillito, V.G., Wing, S.P., Hutton, L.V. e Baker, K.B. *Classification of radar returns from the ionosphere using neural networks*. Johns Hopkins APL Technical Digest, 1989, Vol. 10(3), pp.262-266

Yeh, I.C., Yang, K.J. e Ting, T.M. *Knowledge discovery on RFM model using Bernoulli sequence*. Expert Systems with Applications, 2009, Vol. 36(3), pp.5866-5871

Zhang, S., Qin, Z., Ling, C.X. e Sheng, S. *"Missing is useful": missing values in cost-sensitive decision trees*. IEEE transactions on knowledge and data engineering, 2005, Vol. 17(12), pp.1689-1693