Modelagem Random Forest – Treinamento de Máquinas para indentificação de regiões mais propícias a propagação de COVID-19

**Guilherme Pires Silva1 , Renan Nunes Avila2 1Universidade Tecnológica Federal do Paraná - UTFPR/PR**

[**guisil.2020@alunos.utfpr.edu.br,**](mailto:guisil.2020@alunos.utfpr.edu.br) [**renanavila@alunos.utfpr.edu.br**](mailto:renanavila@alunos.utfpr.edu.br)

1. **Objetivo**

Como objetivo principal com essa modelagem visamos treinar máquinas para a indentificação de áreas mais propícias para a propagação de COVID-19. Com a utilização de uma base de dados construída a partir de dados obtidos pelo IBGE e outros simulados, montamos uma lista de todos 75 bairros registrados em Curitiba, seguidos por suas respectivas média de renda da população que ocupa determinada área, nível de saneamento do bairro e porcentagem de indíviduos com plano de saúde.

1. **Análise da Base de Dados**

Como variáveis independentes temos:

* Bairro;

Como variáveis dependentes temos:

* Niv\_saneam;
* S\_Plan;
* Med\_Salarial;

Como váriavel de saída:

* Risco\_propag;

Com tais informações e os 75 bairros registrados em nossa base de dados, pretendemos treinar máquinas para indentificar regiões mais vulneráveis a propagação de COVID-19. A partir da *tabela 1* analisamos os valores possíveis para cada váriavel.

|  |  |  |
| --- | --- | --- |
| Nome do Atributo | Descrição | Valores Possíveis |
| Bairro | Nome do bairro. | Contínuo |
| Niv\_saneam (%) | Nível de saneamento no bairro. | 0-100 |
| S\_Plan (%) | Porcentagem de indíviduos com plano de saúde no bairro. | 0-100 |
| Med\_Salarial (R$) | Média salarial dos moradores do bairro. | Contínuo |
| Risco\_propag | Vulnerabilidade de maior propagação de COVID-19. | Baixo,Alto |

Tabela 1: Descrição dos atributos da base de dados.

1. **Metodologia**

Com a aplicação do algoritmo de Random Forest, compararemos árvores de decisões CHAID e por fim será votado através do algoritmo a classificação da vulnerabilidade de propagação de COVID-19 para cada um dos 75 bairros registrados.

Árvores a serem analisadas:

1. **Bairro x Niv\_saneam**
2. **Niv\_saneam x Med\_Salarial**
3. **Bairro x S\_Plan**
4. **Bairro x Med\_Salarial**

Modelagem de dados:

* **Nível de saneamento (Niv\_saneam):**

- Abaixo de 55% devemos considerar uma área de ‘Alto’ risco de propagação;

- Acima de 80% devemos considerar uma área de ‘Baixo’ risco de propagação.

* **Média Salarial (Med\_Salarial):**

Nota: Como os dados salariais obtidos são de 2010, onde o sálario minímo era de aproximadamente R$510,00 em comparação a 2021 que é de R$1.192,40. Notamos um aumento de 134% no salario minímo, portanto seguiremos uma média de corte proporcional para simular os anos atuais.

- Abaixo de R$850,00 devemos considerar uma área de ‘Alto’ risco de propagação;

- Acima de R$1600,00 devemos considerar uma área de ‘Baixo’ risco de propagação.

* **Porcentagem de Indíviduos com Plano de Saúde (S\_Plan):**

- Abaixo de 55% devemos considerar uma área de ‘Alto’ risco de propagação;

- Acima de 75% devemos considerar uma área de ‘Baixo’ risco de propagação.

Por fim analisaremos a matriz de confusão e o resultado do treinamento, para garantir que não houve nenhum overfiting e analisar os resultados propostos por essa modelagem.

# Referências

* Pesquisa IBGE 2010 Lista dos Bairros de Curitiba e média de renda, disponíveis no site Wikipédia, em:

<https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Curitiba>