

Faculdade de Ciência e Tecnologia de Montes Claros

Felipe Israel Corrêa

**SISTEMA AGREGADOR PARA ANÁLISE DE
CARACTERÍSTICAS DE IMÓVEIS COM RASTREAMENTO
WEB E APRENDIZADO DE MÁQUINA**

Montes Claros- MG
2018

Felipe Israel Corrêa

**SISTEMA AGREGADOR PARA ANÁLISE DE CARACTERÍSTICAS DE
IMÓVEIS COM RASTREAMENTO WEB E APRENDIZADO DE
MÁQUINA**

Monografia apresentada ao Curso de Engenharia da Computação, da Faculdade de Ciência e Tecnologia de Montes Claros, como parte dos requisitos para obtenção do título de Engenheiro da Computação.

Orientador: **PROF. DR. RENATO DOURADO
MAIA.**

Montes Claros - MG
2018

FUNDAÇÃO EDUCACIONAL MONTES CLAROS
Faculdade de Ciência e Tecnologia de Montes Claros

Felipe Israel Corrêa

**SISTEMA AGREGADOR PARA ANÁLISE DE CARACTERÍSTICAS DE
IMÓVEIS COM RASTREAMENTO WEB E APRENDIZADO DE
MÁQUINA**

Esta monografia foi julgada adequada como parte dos requisitos para a obtenção do diploma de Engenheiro da Computação aprovada pela banca examinadora da Faculdade de Ciência e Tecnologia de Montes Claros.

Prof. Dr. Maurílio José Inácio
Coord. do Curso de Engenharia da Computação

Banca Examinadora

Prof. Dr. Renato Dourado Maia, FACIT / _____
(Orientador)

Prof. Examinador 1 _____

Prof. Examinador 2 _____

Montes Claros, xx de junho de 2018.

Para Lavínia, amor maior.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela saúde.

Aos meus pais Aroldo e Isolina (*In memoriam*) pelos contínuos esforços em me oferecerem uma educação satisfatória.

A minha esposa Laís pelo companheirismo e apoio em toda esta jornada.

Meus irmãos e sobrinhos pela amizade sempre presente.

Aos meus colegas e professores, em especial ao professor Renato Dourado, que como eu, acreditou neste projeto.

“Não se espante com a altura do voo. Quanto mais alto, mais longe do perigo. Quanto mais você se eleva, mais tempo há de reconhecer uma pane. É quando se está próximo do solo que se deve desconfiar. ”

Santos Dumont

RESUMO

A democratização do acesso a internet tem proporcionado uma geração de dados, sobretudo digitais, nunca antes vista na história da humanidade. Nunca se produziu tanta informação em tão pouco tempo. A partir desta perspectiva este estudo tem por objetivo coletar os dados sobre imóveis residenciais, da cidade de Montes Claros, adquiridos em sites de empresas imobiliárias e desenvolver um sistema que os agregue e gere observações relevantes sobre as características destes empreendimentos. A aquisição das informações será feita por meio de um Rastreador Web e a análise dos dados ocorrerá através da aplicação de Aprendizagem de Máquina no que tange a utilização de um algoritmo regressor e um de recomendação e busca. Para alcance do objetivo foram estudados os conceitos que englobam os processos de rastreamento e aprendizagem de máquina, verificadas as condições para desenvolvimento do trabalho e descritos os algoritmos de codificação. Como proposto, o sistema obteve êxito em prever valores de um imóvel com base em suas características, tais como bairro, número de quartos, banheiros, vagas de estacionamento e tamanho da área, além de efetuar a recomendação e busca de imóveis. Todo ele foi confeccionado através da linguagem de programação *Python* no ambiente de desenvolvimento *Visual Studio Code*, em conjunto com as bibliotecas *Requests* e *BeautifulSoup* para o rastreador e as bibliotecas científicas e gráficas *Scikit-learn*, *Pandas*, *Numpy*, *Matplotlib* e *Bokeh* para aprendizagem de máquina.

Palavras-Chave: *Dados, Imóveis, Rastreador Web, Aprendizagem de Máquina, Python.*

ABSTRACT

The popularization of Internet access has provided a generation of data, especially digital data, never before seen in the history of mankind. Never has so much information been produced in such a short time. From this perspective, this study aims to collect data on real estate acquired on websites of real estate companies and develop a system that aggregates and generates relevant observations about the characteristics of these developments. The information acquisition will be done through a Web Tracker and data analysis will take place through the Machine Learning application. In order to reach the objective, the concepts that comprise the processes of machine tracking and learning have been studied, the conditions for the development of the work were verified and the coding algorithms were described. As proposed the system is able to predict values of a property based on its characteristics, such as neighborhood, number of rooms, bathrooms, parking spaces and size of the area, in addition to recommending and searching for real estate. The system was made using the Python programming language in the Visual Studio Code development environment, together with the Requests and BeautifulSoup libraries for the tracker, and the Scikit-learn, Pandas, Numpy, Matplotlib and Bokeh scientific and graphic libraries for learning machine. At the end, this work was successful in developing the system.

Keywords: *Data, Real Estate, Web Tracker, Machine Learning, Python.*

LISTA DE FIGURAS

FIGURA 1 - <i>Document Object Module</i>	17
FIGURA 2 - Código fonte e <i>tags</i> HTML.....	18
FIGURA 3 - Diagrama de fluxo de um rastreador <i>web</i>	18
FIGURA 4 - Fluxo de operações de um sistema de AM.....	23
FIGURA 5 - Representação intercepto e coeficiente de inclinação.....	24
FIGURA 6 - Representação do erro.....	25
FIGURA 7 - Linha de regressão da relação entre variáveis A e B.....	25
FIGURA 8 - Árvore de decisão.....	26
FIGURA 9 - Árvore de decisão e divisões no espaço.....	27
FIGURA 10 - Dissimilaridade calculada entre vetores a e b.....	31
FIGURA 11 - Produto escalar entre os vetores a e b.....	32
FIGURA 12 - Vetores a e b distantes 90°.....	32
FIGURA 13 - Vetores a e b com ângulo igual a 0°.....	32
FIGURA 14 - K-vizinhos mais próximos em classificação.....	33
FIGURA 15 - Diagrama do interpretador <i>Python</i>	35
FIGURA 16 - Exemplos de gráficos produzidos com biblioteca <i>Matplotlib</i>	39
FIGURA 17 - Exemplos de gráficos criados com a biblioteca <i>Bokeh</i>	40
FIGURA 18 - Fluxograma das aplicações.....	41
FIGURA 19 - Script rastreador <i>web</i>	42
FIGURA 20 - Exemplo de busca de <i>tags</i> para rastreador <i>web</i>	43
FIGURA 21 - Exemplo de definição das <i>tags</i> para busca dos dados.....	43
FIGURA 22 - Código para armazenamento dos dados.....	44
FIGURA 23 - Estrutura de criação da tabela no servidor <i>MySQL</i>	44
FIGURA 24 - Códigos para ajuste dos nomes dos bairros.....	46
FIGURA 25 - Porcentagem de valores nulos por coluna.....	46
FIGURA 26 - Código para imputação de valores da média.....	47
FIGURA 27 - Variáveis independentes no hiperplano.....	48
FIGURA 28 - Matriz de correlação entre as variáveis.....	48
FIGURA 29 - Processo de validação cruzada.....	49
FIGURA 30 - Separação do conjunto de dados entre treino e teste.....	50
FIGURA 31 - Método de avaliação do modelo linear.....	50
FIGURA 32 - Construção do modelo.....	51

FIGURA 33 - Avaliação e definição da profundidade da árvore de decisão.....	52
FIGURA 34 - Teste do modelo com 1 a 100 árvores.....	52
FIGURA 35 - Cálculo para floresta aleatória.....	53
FIGURA 36 - Erro global entre valores reais e preditos.....	55
FIGURA 37 - Média das características dos imóveis.....	56
FIGURA 38 - Predição de preço do imóvel.....	57
FIGURA 39 - Preços para apartamentos de 1 quarto.....	57
FIGURA 40 - Preços para apartamentos de 2 quartos.....	58
FIGURA 41 - Preços para apartamentos de 3 quartos.....	58
FIGURA 42 - Preços para apartamentos de 4 quartos.....	59
FIGURA 43 - Preços para apartamentos de 5 quartos.....	59
FIGURA 44 - Busca e armazenamento das coordenadas geográficas.....	61
FIGURA 45 - Leitura e atribuição dos conjuntos de dados.....	62
FIGURA 46 - Inserção das coordenadas geográficas.....	62
FIGURA 47 - Calculador de dissimilaridade.....	62
FIGURA 48 - Selecionar imóveis de acordo com o bairro.....	63
FIGURA 49 - Busca e recomendação de apartamentos.....	63
FIGURA 50 - Chamada dos métodos de busca e recomendação.....	63
FIGURA 51 - Resultado busca e recomendação para métrica euclidiana.....	64
FIGURA 52 - Resultado busca e recomendação para métrica do cosseno.....	64
FIGURA 53 - Itens recomendados e valor da dissimilaridade.....	65
FIGURA 54 - Implementação para cálculo da precisão e <i>recall</i>	66
FIGURA 55 - Implementação fórmula <i>F1-score</i>	67
FIGURA 56 - Imóveis apresentados calculados pela dissimilaridade euclidiana.....	67
FIGURA 57 - Imóveis apresentados calculados pela similaridade do cosseno.....	67
FIGURA 58 - Quantidade de imóveis por bairro.....	69
FIGURA 59 - Distribuição de imóveis pela cidade.....	70
FIGURA 60 - Quantidade de imóveis, por bairro, com 1 e 2 quartos.....	70
FIGURA 61 - Quantidade de imóveis, por bairro, com 3 e 4 quartos.....	71
FIGURA 62 - Quantidade de imóveis, por bairro, com 5 quartos.....	71
FIGURA 63 - Porcentagem de banheiros e vagas de garagem.....	72
FIGURA 64 – Estimativa de valor para o menor apartamento.....	72
FIGURA 65 - Estimativa de valor para o maior apartamento.....	73
FIGURA 66 - Estimativa de valor com área aumentada em 10m ²	73

FIGURA 67 - Diferença no valor do preço do imóvel a cada aumento da área.....	73
FIGURA 68 - Diferença de preços baseada no aumento do número de quartos.....	74
FIGURA 69 - Diferença de preços baseada no aumento do número de banheiros...	74
FIGURA 70 - Diferença de preços baseada no aumento do número de vagas.....	75

LISTA DE TABELAS

TABELA 1 - Característica dos imóveis.....	45
TABELA 2 - Conjunto de dados após pré-processamento.....	47
TABELA 3 - Resultado do cálculo das métricas do modelo linear.....	51
TABELA 4 - Resultado das métricas da árvore de decisão de acordo com sua profundidade.....	53
TABELA 5 - Resultado das métricas da floresta aleatória de acordo com quantidade de árvores e profundidade	54
TABELA 6 - Média das características dos apartamentos com base na quantidade de quartos.....	56
TABELA 7 - Resultado cálculos de eficiência das recomendações	68

SUMÁRIO

INTRODUÇÃO.....	14
CAPÍTULO 1 SISTEMA AGREGADOR PARA ANÁLISE DE CARACTERÍSTICAS DE IMÓVEIS COM RASTREAMENTO WEB E APRENDIZADO DE MÁQUINA.....	16
1.1 Rastreador Web.....	16
1.2 Fluxo de Busca.....	17
1.3 Aprendizado de Máquina.....	19
1.3.1 Aplicações.....	20
1.4 Tarefas de Aprendizado.....	20
1.5 Tipos de Aprendizagem.....	21
1.5.1 Aprendizagem Supervisionada.....	21
1.5.2 Aprendizagem Não Supervisionada.....	22
1.6 Fluxo de Funcionamento de um Sistema AM.....	22
1.7 Sistema de Regressão.....	23
1.7.1 Regressão Linear Simples e Múltipla.....	23
1.7.2 Árvore de Decisão e Floresta Aleatória.....	26
1.8 Sistema de Recomendação.....	29
1.9 Tipos Sistema de Recomendação.....	29
1.9.1 Filtragem Colaborativa.....	29
1.9.2 Recomendação Baseada em Conteúdo.....	30
1.10 Técnicas de Recomendação.....	30
1.10.1 Recomendação Baseada em Vizinhaça.....	30
1.11 Distância Euclidiana.....	31
1.12 Similaridade do Cosseno.....	31
1.13 K-Vizinhos Mais Próximos.....	33
CAPÍTULO 2 MATERIAIS E MÉTODOS.....	34
2.1 Ambiente de Desenvolvimento.....	34
2.2 Linguagem Python.....	35
2.3 Bibliotecas.....	36
2.3.1 Requests.....	36
2.3.2 BeautifulSoup.....	37
2.3.3 Googlemaps Geocoding.....	37

2.3.4 Mysql.connector.....	37
2.3.5 Scikit-learn.....	37
2.3.6 Numpy.....	38
2.3.7 Pandas.....	38
2.3.8 Matplotlib.....	39
2.3.9 Bokeh.....	39
2.4 Questionário.....	40
CAPÍTULO 3 RESULTADOS: APRESENTAÇÃO, ANÁLISE E DISCUSSÃO.....	41
3.1 Rastreador Web.....	41
3.2 Regressor.....	45
3.3 Sistema de Recomendação e Busca.....	60
3.4 Análise Exploratória dos Dados.....	68
CAPÍTULO 4 APLICAÇÃO.....	76
CONSIDERAÇÕES FINAIS.....	77
REFERÊNCIAS.....	79