Big Data Course

Capstone Project   
Action Plan

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |  |
| --- | --- |
| Course | Big Data Course |
| Team Name | Nhóm 1 – BD02 |
| Team Leader/  Members | Lê Đông Anh Kiệt /  Trần Tuấn Kiệt, Ngô Văn Lâu, Ngân Hoàng Huy, La Hồng Lộc, Nguyễn Đặng Phú Mẫn |
| Project Title | Phân tích dự đoán giá nhà đất |
| Goal |  |
| Thu thập dữ liệu giá nhà đất ở Bắc Kinh và thông qua quá trình tiền xử lý, phân tích và đánh giá dữ liệu từ đó chọn các mô hình thuật toán thích hợp để huấn luyện sau đó đánh giá độ hiểu quả từ đó chọn mô hình tốt nhất cho bài toán để thực hiện việc triển khai quá trình dự đoán giá nhà đất ở Bắc Kinh. | |
| Abstract |  |
| Trong những năm gần đây, thị trường bất động sản Bắc Kinh có những biến động lớn về giá cả, ảnh hưởng không nhỏ đến nền kinh tế xã hội. Dự án này nhằm phát triển một mô hình phân tích dự đoán giá nhà đất tại Bắc Kinh, sử dụng các thuật toán máy học để phân tích dữ liệu từ các giao dịch bất động sản và cung cấp cái nhìn sâu sắc hơn về xu hướng giá thành và hỗ trợ các nhà đầu tư, người tiêu dùng trong việc đưa ra quyết định chính xác nhất. Để hoàn thành dự án nhóm đã thu thập và tiền xử lý dữ liệu như xử lý thông tin, xử lý cột, biến đổi cột từ đó được bộ dữ liệu sạch. Tiếp theo nhóm đã phân tích dữ liệu thông qua thống kê từ đó lựa chọn mô hình Linear Regression, Decision Tree, Random Forest, Gradient Bootsting và Neural Network. Sau khi nhóm huấn luyện mô hình xong đã đánh giá mô hình và lựa chọn mô hình dự đoán tốt nhất, đó là mô hình Random Forest. | |

|  |  |
| --- | --- |
| Method |  |
| Tiền xử lý dữ liệu như như xử lý thông tin, xử lý cột, biến đổi cột từ đó được bộ dữ liệu sạch sử dụng kĩ thuật như Min-Max Scaler, xử lý giá trị ngoại lệ bằng Interquartile Range.  Phân tích dữ liệu thông qua thống kê.  Lựa chọn mô hình Neuron Network, Linear Regression, Decision Tree, Random Forest, Gradient Bootsting, MLP sau huấn luyện mô hình.  Đánh giá mô hình và lựa chọn mô hình hiệu quả nhất. | |
| Data |  |
| Dữ liệu được sử dụng trong dự án này bao gồm 318,851 bản ghi giao dịch bất động sản từ Bắc Kinh, Trung Quốc. Những bản ghi này bao gồm nhiều thuộc tính khác nhau cung cấp những thông tin quan trọng về động thái của thị trường bất động sản ở Bắc Kinh. Mỗi thuộc tính đều đóng vai trò quan trọng trong việc hiểu rõ giá bất động sản và xu hướng thị trường.  Bộ dữ liệu bao gồm 26 thuộc tính, bao gồm các thông tin cơ bản như ID giao dịch, tọa độ vị trí, ngày giao dịch, giá bất động sản, diện tích, chi tiết phòng, loại tòa nhà, thời gian xây dựng, điều kiện sửa chữa, tiện nghi, thông tin quận huyện, và giá trung bình của cộng đồng.  Mục tiêu chính của việc lựa chọn bộ dữ liệu này là phát triển một mô hình dự đoán giá bất động sản ở Bắc Kinh. Bằng cách sử dụng các thuật toán học máy và kỹ thuật phân tích dữ liệu, mục tiêu là cung cấp hiểu biết chi tiết về xu hướng giá, hỗ trợ ra quyết định cho nhà đầu tư và người tiêu dùng, và nâng cao sự hiểu biết tổng thể về thị trường bất động sản ở Bắc Kinh.  Trước khi phân tích, bộ dữ liệu đã trải qua các bước tiền xử lý để giải quyết các giá trị thiếu, xử lý ngoại lệ, và đảm bảo chất lượng dữ liệu. Các quy trình làm sạch đã được thực hiện để chuẩn bị dữ liệu cho việc mô hình hóa, đảm bảo rằng nó chính xác và phù hợp cho phân tích.  Bộ dữ liệu được chọn là quan trọng để đánh giá các yếu tố ảnh hưởng đến giá bất động sản ở Bắc Kinh. Bằng cách xem xét các thuộc tính như vị trí, đặc điểm bất động sản, tiện nghi, và xu hướng thị trường, các bên liên quan có thể thu được những thông tin giá trị giúp hỗ trợ ra quyết định chiến lược và phân tích thị trường. | |
| Expected  Outcome |  |
| Nhóm mong muốn xử lý tốt đữ liệu đã thu thập từ đó có được bộ dữ liệu sạch để huấn luyện các mô hình một cách chính xác.  Đồng thời lựa chọn các mô hình thích hợp để huấn luyện từ đó có thể so sánh, đánh giá các mô hình.  Từ đó các đánh giá đó nhóm sẽ lựa chọn mô hình tốt nhất để áp dụng để dự doán giá nhà đất ở Bắc Kinh. | |

|  |  |
| --- | --- |
| Role by  Member |  |
| * Lê Đông Anh Kiệt: Nhóm trưởng * Tiền xử lý dữ liệu * Hỗ trợ phân tích, trực quan hóa dữ liệu * Huấn luyện mô hình * Trần Tuấn Kiệt: * Hỗ trợ phân tích, trực quan hóa dữ liệu * Hỗ trợ huấn luyện, đánh giá mô hình * Báo cáo thuyết trình * Ngô Văn Lâu: * Báo cáo dự án * Hỗ trợ thu thập dữ liệu * Đánh giá mô hình * Ngân Hoàng Huy: * Hỗ trợ báo cáo dự án * Hỗ trợ thu thập dữ liệu * Phân tích, đánh giá mô hình * La Hồng Lộc: * Hỗ trợ báo cáo dự án * Hỗ trợ thu thập dữ liệu * Nguyễn Đặng Phú Mẫn: * Hỗ trợ báo cáo thuyết trình * Hỗ trợ phân tích, trực quan hóa dữ liệu | |

|  |  |
| --- | --- |
| Schedule  Summary |  |
| 1/6/2024 – 24/6/2024: Thu thập dữ liệu   * 1/6/2024 - 10/06/2024 : Thu thập dữ liệu * 11/6/2024 - 24/6/2024: Chuẩn bị dữ liệu   18/6/2024 - 07/07/2024: Tiền xử lý dữ liệu   * 18/6/2024 - 25/6/2024: Làm sạch dữ liệu * 26/6/2024 - 03/07/2024: Trích lọc đặc trưng * 04/07/2024 - 07/07/2024: Phân chia dữ liệu   08/07/2024 - 14/8/2024: Phát triển mô hình   * 08/07/2024 - 10/07/2024: Lựa chọn mô hình * 11/07/2024 - 24/7/2024: Huấn luyện mô hình * 09/08/2024 - 14/8/2024: Lựa chọn mô hình tốt nhất | |
| Comment &  Assessment |  |
| <Comment and assessment **by the instructor.**> | |