TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT

KHOA: CÔNG NGHỆ THÔNG TIN

A blue circle with white text and a globe

Description automatically generated

ĐỒ ÁN MÔN HỌC

***Họ và tên: Hoàng Hồng Huy***

***Mã sinh viên: 2121050316***

***Lớp: DCCTCT66\_09D***

***Học phần: 7080205 – Các hệ thống thông tin thông minh – nhóm 200***

Mục lục

[1](#_Toc184894232)

[LỜI MỞ ĐẦU 3](#_Toc184894233)

[Chương 1: Mô tả bài toán 4](#_Toc184894234)

[1.1 Mục tiêu của bài toán 4](#_Toc184894235)

[1.1.1 Làm gì? 4](#_Toc184894236)

[1.1.2 Cho đối tượng nào? 5](#_Toc184894237)

[1.2. Mô tả các yêu cầu đặt ra của bài toán 6](#_Toc184894238)

[1.2.1. Yêu cầu về dữ liệu 6](#_Toc184894239)

[1.2.2. Yêu cầu về chức năng xử lý 6](#_Toc184894240)

[1.2.3. Yêu cầu về quy trình xử lý 7](#_Toc184894241)

[1.3. Mô tả các nguồn tài liệu, dữ liệu, mẫu biểu liên quan đến bài toán 7](#_Toc184894242)

[1.3.1. Tài liệu 7](#_Toc184894243)

[1.3.2. Dữ liệu 7](#_Toc184894244)

[1.3.3. Mẫu biểu 8](#_Toc184894245)

[1.4. Phân tích, đánh giá hiện trạng và những vấn đề còn tồn tại 8](#_Toc184894246)

[1.4.1 Hiện trạng quản lý cây xanh đô thị 8](#_Toc184894247)

[1.4.2 Những vấn đề còn tồn tại 8](#_Toc184894248)

[1.4.3 Những vấn đề mà đề tài cần tập trung nghiên cứu giải quyết 9](#_Toc184894249)

[Chương 2: Cơ sở lý thuyết và cơ sở thực nghiệm về vấn đề cần giải quyết 9](#_Toc184894250)

[2.1. Cơ sở lý thuyết 9](#_Toc184894251)

[2.1.1. Lý thuyết về cây quyết định 9](#_Toc184894252)

[2.1.2. Entropy và Information Gain 10](#_Toc184894253)

[2.1.3. Mô hình học máy và trí tuệ nhân tạo 10](#_Toc184894254)

[2.2. Giả thuyết khoa học 11](#_Toc184894255)

[2.3. Phương pháp kỹ thuật và công nghệ 11](#_Toc184894256)

[2.3.1. Ngôn ngữ lập trình và công nghệ web 11](#_Toc184894257)

[2.3.2. Thư viện và công cụ hỗ trợ 11](#_Toc184894258)

[2.3.3. Quy trình thực hiện đề tài 11](#_Toc184894259)

[2.4. Đánh giá, nhận định tại sao lại chọn công nghệ/ngôn ngữ lập trình/framework đó để thực hiện đề tài 12](#_Toc184894260)

[Chương 3: Giải quyết bài toán 15](#_Toc184894261)

[3.1. Phân tích các yêu cầu 15](#_Toc184894262)

[3.1.1. Yêu cầu về dữ liệu 16](#_Toc184894263)

[3.1.2. Yêu cầu về chức năng xử lý 16](#_Toc184894264)

[3.1.3. Yêu cầu về quy trình xử lý 16](#_Toc184894265)

[3.2. Thiết kế hệ thống 16](#_Toc184894266)

[3.2.1. Kiến trúc hệ thống 16](#_Toc184894267)

[3.2.2. Thiết kế cơ sở dữ liệu 17](#_Toc184894268)

[3.2.3. Thiết kế giao diện người dùng 17](#_Toc184894269)

[3.3. Cài đặt hệ thống 18](#_Toc184894270)

[3.3.1. Thiết lập môi trường phát triển 18](#_Toc184894271)

[3.3.2. Cài đặt các thư viện và công cụ hỗ trợ 18](#_Toc184894272)

[3.3.3. Triển khai các thành phần của hệ thống 19](#_Toc184894273)

[Chương 4: Kết quả đạt được 19](#_Toc184894274)

[4.1. Thu thập và xử lý dữ liệu 19](#_Toc184894275)

[4.2. Huấn luyện mô hình cây quyết định 21](#_Toc184894276)

[4.3. Phân tích và dự đoán tình trạng sức khỏe cây 28](#_Toc184894277)

[4.4. Đánh giá kết quả đạt được 28](#_Toc184894278)

[Kết luận 29](#_Toc184894279)

# LỜI MỞ ĐẦU

**Lý do chọn đề tài**

Môi trường đô thị ngày càng phát triển, kéo theo nhu cầu về không gian xanh, cây xanh đô thị ngày càng trở nên cấp thiết. Cây xanh không chỉ mang lại vẻ đẹp thẩm mỹ mà còn đóng vai trò quan trọng trong việc cải thiện chất lượng không khí, giảm ô nhiễm môi trường và cung cấp bóng mát cho con người. Bên cạnh đó, cây xanh còn góp phần tạo nên cảnh quan thiên nhiên hài hòa, giúp giảm căng thẳng, tăng cường sức khỏe tinh thần và thể chất cho người dân.

Tuy nhiên, việc quản lý và duy trì cây xanh đô thị gặp không ít khó khăn do số lượng cây lớn, phân bố rải rác khắp nơi và đòi hỏi sự theo dõi, chăm sóc liên tục. Các vấn đề như sâu bệnh, gãy đổ, cây chết cần được phát hiện và xử lý kịp thời để đảm bảo an toàn và mỹ quan đô thị. Do đó, việc ứng dụng công nghệ thông tin vào quản lý cây xanh là một xu hướng tất yếu và cần thiết trong bối cảnh hiện nay.

**Mục tiêu của đề tài**

Đề tài này nhằm mục tiêu xây dựng một hệ thống quản lý cây xanh đô thị hiệu quả, áp dụng các phương pháp khoa học hiện đại để giám sát và phân tích tình trạng sức khỏe của cây. Cụ thể, hệ thống sẽ hỗ trợ:

* Theo dõi và cập nhật thông tin chi tiết về cây xanh như loại cây, chiều cao, đường kính, tuổi, tình trạng sức khỏe, vị trí, ghi chú.
* Dự đoán tình trạng sức khỏe của cây dựa trên các đặc điểm vật lý và các dữ liệu lịch sử.
* Tạo ra các báo cáo, biểu đồ trực quan giúp người quản lý dễ dàng theo dõi và ra quyết định.
* Cảnh báo sớm về các vấn đề sức khỏe của cây để có biện pháp xử lý kịp thời.

**Đối tượng và phạm vi của đề tài**

* **Đối tượng nghiên cứu**: Đối tượng của đề tài là các cây xanh đô thị, bao gồm các loài cây được trồng trong khu vực thành phố.
* **Phạm vi nghiên cứu**: Hệ thống quản lý sẽ bao gồm các chức năng từ theo dõi, cập nhật thông tin cây xanh, dự đoán tình trạng sức khỏe, đến việc tạo ra các báo cáo, biểu đồ phân tích và cảnh báo sớm. Đề tài sẽ giới hạn trong phạm vi nghiên cứu cây xanh tại một thành phố cụ thể, nhưng có thể mở rộng và áp dụng cho nhiều khu vực khác trong tương lai.

Trong quá trình nghiên cứu và triển khai đề tài, các phương pháp khoa học, công nghệ hiện đại như trí tuệ nhân tạo, học máy và các thuật toán phân tích dữ liệu sẽ được áp dụng. Hệ thống sẽ sử dụng ngôn ngữ lập trình Python và các thư viện hỗ trợ như Pandas, Scikit-learn, Plotly và Streamlit để xây dựng các chức năng và giao diện người dùng thân thiện, dễ sử dụng.

# Chương 1: Mô tả bài toán

## **1.1 Mục tiêu của bài toán**

### **1.1.1 Làm gì?**

Bài toán đặt ra là xây dựng một hệ thống quản lý cây xanh đô thị hiệu quả, nhằm giúp giám sát, cập nhật thông tin và dự đoán tình trạng sức khỏe của cây xanh một cách chính xác và kịp thời. Việc quản lý cây xanh đô thị không chỉ đơn thuần là việc trồng và chăm sóc cây mà còn bao gồm rất nhiều hoạt động phức tạp khác như theo dõi sự phát triển của cây, phát hiện sớm các vấn đề sức khỏe, đưa ra các biện pháp chăm sóc và bảo dưỡng phù hợp. Một hệ thống quản lý cây xanh hiện đại sẽ giúp tối ưu hóa các hoạt động này, đảm bảo rằng cây xanh đô thị luôn trong tình trạng tốt nhất, từ đó góp phần tạo ra một môi trường sống trong lành, an toàn và đẹp mắt cho người dân.

Hệ thống sẽ bao gồm các chức năng chính sau:

1. **Giám sát cây xanh**: Hệ thống sẽ cung cấp các công cụ để theo dõi tình trạng sức khỏe của cây xanh theo thời gian. Các dữ liệu về chiều cao, đường kính, tuổi, loài cây, vị trí, và tình trạng sức khỏe của cây sẽ được cập nhật và lưu trữ trong cơ sở dữ liệu. Điều này giúp các nhà quản lý có cái nhìn tổng quan về tình trạng cây xanh trong khu vực đô thị.
2. **Cập nhật thông tin**: Người dùng có thể cập nhật thông tin về cây xanh một cách dễ dàng. Hệ thống sẽ hỗ trợ nhập liệu các thông tin mới như cây mới trồng, các biện pháp chăm sóc và bảo dưỡng đã thực hiện, các sự kiện quan trọng liên quan đến cây. Các thông tin này sẽ giúp hệ thống duy trì dữ liệu chính xác và kịp thời.
3. **Dự đoán tình trạng sức khỏe**: Sử dụng các thuật toán học máy, hệ thống sẽ có khả năng dự đoán tình trạng sức khỏe của cây dựa trên các đặc điểm vật lý và dữ liệu lịch sử. Việc dự đoán này giúp phát hiện sớm các vấn đề sức khỏe của cây, từ đó đưa ra các biện pháp phòng ngừa hoặc khắc phục kịp thời, tránh tình trạng cây bị hỏng hóc hoặc gây nguy hiểm cho người dân.
4. **Tạo báo cáo và biểu đồ phân tích**: Hệ thống sẽ cung cấp các báo cáo và biểu đồ phân tích trực quan, giúp người dùng dễ dàng theo dõi và đánh giá tình trạng sức khỏe của cây xanh. Các báo cáo này có thể bao gồm các thông tin như số lượng cây theo tình trạng sức khỏe, phân bố chiều cao và đường kính cây, tỷ lệ các loài cây, v.v. Điều này giúp người quản lý có cái nhìn tổng quan và chi tiết về tình trạng cây xanh, từ đó đưa ra các quyết định quản lý chính xác và hiệu quả.

Việc xây dựng một hệ thống quản lý cây xanh đô thị hiện đại không chỉ giúp cải thiện hiệu quả quản lý mà còn góp phần quan trọng vào việc bảo vệ môi trường và nâng cao chất lượng cuộc sống của người dân. Hệ thống này sẽ giúp các cơ quan quản lý đô thị và các tổ chức môi trường dễ dàng theo dõi và quản lý cây xanh, từ đó đảm bảo rằng cây xanh đô thị luôn trong tình trạng tốt nhất và đóng góp tích cực vào môi trường sống của con người.

### **1.1.2 Cho đối tượng nào?**

Đối tượng sử dụng hệ thống quản lý cây xanh đô thị bao gồm các cơ quan quản lý đô thị, các tổ chức môi trường, các công ty dịch vụ cây xanh, và cả người dân. Mỗi đối tượng có những nhu cầu và yêu cầu riêng đối với hệ thống, cụ thể:

1. **Các cơ quan quản lý đô thị**:
   * **Nhu cầu**: Cần có một hệ thống giám sát và quản lý cây xanh hiệu quả để đảm bảo an toàn cho người dân, duy trì mỹ quan đô thị và bảo vệ môi trường. Các cơ quan này cần theo dõi tình trạng sức khỏe của cây, phát hiện sớm các vấn đề và có biện pháp xử lý kịp thời.
   * **Yêu cầu**: Hệ thống phải cung cấp các chức năng giám sát, cập nhật thông tin, dự đoán tình trạng sức khỏe, tạo báo cáo và biểu đồ phân tích, cảnh báo sớm và phân quyền quản lý.
2. **Các tổ chức môi trường**:
   * **Nhu cầu**: Cần thu thập và phân tích dữ liệu về cây xanh để nghiên cứu và bảo vệ môi trường. Các tổ chức này cần thông tin chi tiết về tình trạng sức khỏe của cây, các biện pháp chăm sóc và bảo dưỡng, và các yếu tố ảnh hưởng đến sức khỏe của cây.
   * **Yêu cầu**: Hệ thống phải cung cấp các dữ liệu chi tiết về cây xanh, hỗ trợ phân tích và tạo báo cáo, biểu đồ, và cung cấp các thông tin khoa học liên quan.
3. **Các công ty dịch vụ cây xanh**:
   * **Nhu cầu**: Cần quản lý và theo dõi các dịch vụ chăm sóc và bảo dưỡng cây xanh. Các công ty này cần thông tin về các yêu cầu dịch vụ, tình trạng sức khỏe của cây, và lịch trình chăm sóc và bảo dưỡng.
   * **Yêu cầu**: Hệ thống phải cung cấp các chức năng cập nhật thông tin dịch vụ, theo dõi tình trạng sức khỏe của cây, và quản lý lịch trình chăm sóc và bảo dưỡng.
4. **Người dân**:
   * **Nhu cầu**: Cần có thông tin về cây xanh trong khu vực sống của mình để bảo vệ và duy trì cây xanh. Người dân cần biết về tình trạng sức khỏe của cây, các biện pháp chăm sóc và bảo dưỡng, và có thể tham gia vào các hoạt động bảo vệ cây xanh.
   * **Yêu cầu**: Hệ thống phải cung cấp thông tin chi tiết về cây xanh, hỗ trợ tìm kiếm và tra cứu thông tin, và có chức năng phản hồi và báo cáo về tình trạng cây.

Hệ thống quản lý cây xanh đô thị hiện đại sẽ đáp ứng đầy đủ nhu cầu và yêu cầu của các đối tượng sử dụng, góp phần nâng cao hiệu quả quản lý cây xanh, bảo vệ môi trường và cải thiện chất lượng cuộc sống của người dân. Việc tích hợp các công nghệ hiện đại vào quản lý cây xanh không chỉ mang lại lợi ích kinh tế mà còn tạo ra giá trị bền vững cho cộng đồng và môi trường.

## 1.2. Mô tả các yêu cầu đặt ra của bài toán

### 1.2.1. Yêu cầu về dữ liệu

* **Thông tin cây xanh**: Bao gồm các đặc điểm vật lý như chiều cao, đường kính, tuổi cây, loài cây, tình trạng sức khỏe, vị trí, ghi chú và ngày nhắc nhở.
* **Thông tin về loài cây**: Cung cấp các thông tin cơ bản về từng loài cây như tên loài, đặc điểm sinh học, điều kiện sống, và các thông tin liên quan.

### 1.2.2. Yêu cầu về chức năng xử lý

* **Theo dõi và cập nhật thông tin cây xanh**: Hệ thống cần cho phép người dùng cập nhật thông tin về cây xanh, bao gồm các đặc điểm vật lý và tình trạng sức khỏe hiện tại của cây.
* **Dự đoán tình trạng sức khỏe của cây**: Sử dụng các thuật toán học máy để dự đoán tình trạng sức khỏe của cây dựa trên các đặc điểm vật lý.
* **Tạo báo cáo và biểu đồ phân tích**: Hệ thống cần cung cấp các báo cáo, biểu đồ trực quan giúp người dùng dễ dàng theo dõi và phân tích tình trạng sức khỏe của cây xanh.
* **Cảnh báo sớm về các vấn đề sức khỏe của cây**: Hệ thống cần có chức năng cảnh báo sớm về các vấn đề sức khỏe của cây để người quản lý có thể đưa ra biện pháp xử lý kịp thời.

### 1.2.3. Yêu cầu về quy trình xử lý

* **Thu thập dữ liệu**: Dữ liệu cần được thu thập từ các nguồn đáng tin cậy như cơ quan quản lý đô thị, các tổ chức môi trường và các cơ sở dữ liệu hiện có.
* **Xử lý dữ liệu**: Dữ liệu cần được xử lý và làm sạch trước khi sử dụng trong các thuật toán học máy.
* **Huấn luyện mô hình**: Sử dụng các thuật toán học máy để huấn luyện mô hình dự đoán tình trạng sức khỏe của cây.
* **Kiểm thử và đánh giá**: Hệ thống cần được kiểm thử và đánh giá trên các dữ liệu thực tế để đảm bảo độ chính xác và hiệu quả.

## 1.3. Mô tả các nguồn tài liệu, dữ liệu, mẫu biểu liên quan đến bài toán

### 1.3.1. Tài liệu

* **Tài liệu về các loài cây**: Các sách, báo, tạp chí và các tài liệu khoa học cung cấp thông tin về các loài cây, đặc điểm sinh học, điều kiện sống và các yếu tố ảnh hưởng đến sức khỏe của cây.
* **Tài liệu về học máy và trí tuệ nhân tạo**: Các sách và tài liệu học thuật về các thuật toán học máy, trí tuệ nhân tạo, và các phương pháp phân tích dữ liệu.
* **Tài liệu về quản lý đô thị**: Các tài liệu và báo cáo về quản lý cây xanh đô thị, các chính sách và quy định liên quan.

### 1.3.2. Dữ liệu

* **Dữ liệu cây xanh**: Thông tin về các cây xanh trong khu vực đô thị, bao gồm các đặc điểm vật lý và tình trạng sức khỏe.
* **Dữ liệu lịch sử**: Các dữ liệu lịch sử về tình trạng sức khỏe của cây, các sự kiện chăm sóc, bảo dưỡng, và các thay đổi quan trọng trong quá khứ.
* **Dữ liệu môi trường**: Các thông tin về điều kiện môi trường như khí hậu, đất đai, lượng mưa, và các yếu tố khác ảnh hưởng đến sức khỏe của cây.

### 1.3.3. Mẫu biểu

* **Mẫu biểu thu thập dữ liệu cây xanh**: Mẫu biểu để thu thập thông tin về các cây xanh, bao gồm các trường thông tin như tên loài, chiều cao, đường kính, tuổi, tình trạng sức khỏe, vị trí, ghi chú và ngày nhắc nhở.
* **Mẫu biểu cập nhật thông tin cây xanh**: Mẫu biểu để cập nhật thông tin về tình trạng sức khỏe, các sự kiện chăm sóc và bảo dưỡng, và các thay đổi quan trọng của cây.

## **1.4. Phân tích, đánh giá hiện trạng và những vấn đề còn tồn tại**

### 1.4.1 **Hiện trạng quản lý cây xanh đô thị**

Hiện nay, việc quản lý cây xanh đô thị chủ yếu dựa vào các phương pháp thủ công và giấy tờ, gây khó khăn trong việc theo dõi, cập nhật thông tin và đưa ra các quyết định kịp thời. Các cơ quan quản lý thường gặp khó khăn trong việc giám sát tình trạng sức khỏe của cây, phát hiện sớm các vấn đề và triển khai các biện pháp xử lý hiệu quả.

Các hệ thống quản lý cây xanh hiện tại thường thiếu tính đồng bộ và cập nhật thông tin không chính xác. Các dữ liệu về tình trạng sức khỏe của cây thường không được thu thập và lưu trữ một cách có hệ thống, dẫn đến việc khó khăn trong việc phân tích và ra quyết định. Bên cạnh đó, các công cụ hỗ trợ phân tích và dự đoán tình trạng sức khỏe của cây còn hạn chế, không đủ để đáp ứng nhu cầu quản lý cây xanh một cách hiệu quả.

### 1.4.2 **Những vấn đề còn tồn tại**

**Thiếu thông tin chính xác và cập nhật**: Các thông tin về cây xanh thường không được cập nhật đầy đủ và kịp thời, gây khó khăn trong việc đánh giá tình trạng sức khỏe và triển khai các biện pháp chăm sóc, bảo dưỡng. Việc thiếu thông tin chính xác và cập nhật cũng ảnh hưởng đến khả năng dự đoán và cảnh báo sớm về các vấn đề sức khỏe của cây.

**Khó khăn trong việc dự đoán và cảnh báo sớm**: Việc thiếu các công cụ và phương pháp khoa học để dự đoán tình trạng sức khỏe của cây khiến cho các vấn đề không được phát hiện sớm và xử lý kịp thời. Điều này dẫn đến việc cây xanh có thể bị hỏng hóc, gãy đổ, gây nguy hiểm cho người dân và ảnh hưởng đến mỹ quan đô thị.

**Thiếu sự phối hợp và chia sẻ thông tin**: Các cơ quan, tổ chức liên quan thường hoạt động riêng lẻ, thiếu sự phối hợp và chia sẻ thông tin, dẫn đến việc quản lý cây xanh không hiệu quả. Việc thiếu sự phối hợp này cũng làm giảm khả năng phản ứng nhanh chóng và hiệu quả đối với các vấn đề khẩn cấp liên quan đến cây xanh.

**Thiếu các công cụ phân tích và báo cáo**: Hiện tại, các công cụ hỗ trợ phân tích, báo cáo và hiển thị thông tin về cây xanh còn hạn chế, gây khó khăn trong việc theo dõi và đưa ra các quyết định quản lý. Việc thiếu các công cụ này cũng làm giảm khả năng đánh giá và cải thiện các chính sách quản lý cây xanh.

### 1.4.3 **Những vấn đề mà đề tài cần tập trung nghiên cứu giải quyết**

**Xây dựng hệ thống quản lý cây xanh đô thị hiện đại**: Sử dụng các công nghệ thông tin để xây dựng một hệ thống quản lý cây xanh hiệu quả, giúp theo dõi, cập nhật thông tin, dự đoán tình trạng sức khỏe và đưa ra các cảnh báo sớm về các vấn đề sức khỏe của cây.

**Tích hợp các công cụ phân tích và báo cáo**: Phát triển các công cụ hỗ trợ phân tích, báo cáo và hiển thị thông tin về cây xanh một cách trực quan và dễ hiểu. Các công cụ này cần giúp người quản lý dễ dàng theo dõi tình trạng sức khỏe của cây, đánh giá hiệu quả của các biện pháp chăm sóc và bảo dưỡng, và đưa ra các quyết định kịp thời và chính xác.

**Phát triển các thuật toán dự đoán và cảnh báo sớm**: Sử dụng các thuật toán học máy và trí tuệ nhân tạo để phát triển các mô hình dự đoán tình trạng sức khỏe của cây, dựa trên các đặc điểm vật lý và dữ liệu lịch sử. Các mô hình này cần có khả năng đưa ra các cảnh báo sớm về các vấn đề sức khỏe của cây, giúp người quản lý có thể đưa ra các biện pháp xử lý kịp thời.

**Nâng cao khả năng phối hợp và chia sẻ thông tin**: Phát triển các giải pháp công nghệ giúp nâng cao khả năng phối hợp và chia sẻ thông tin giữa các cơ quan, tổ chức liên quan, đảm bảo tính đồng bộ và chính xác của dữ liệu. Điều này giúp cải thiện hiệu quả quản lý cây xanh và đảm bảo sự phản ứng nhanh chóng và hiệu quả đối với các vấn đề khẩn cấp liên quan đến cây xanh.

# Chương 2: Cơ sở lý thuyết và cơ sở thực nghiệm về vấn đề cần giải quyết

## 2.1. Cơ sở lý thuyết

### 2.1.1. Lý thuyết về cây quyết định

Cây quyết định là một thuật toán học máy thuộc nhóm giám sát (supervised learning), được sử dụng để phân loại và dự đoán. Cây quyết định mô hình hóa các quyết định và các hậu quả của chúng dưới dạng cây phân nhánh. Một cây quyết định bao gồm các thành phần chính như nút gốc, nút trung gian và nút lá:

* **Nút gốc (Root Node)**: Nút bắt đầu của cây quyết định, đại diện cho toàn bộ tập dữ liệu.
* **Nút trung gian (Internal Nodes)**: Các nút bên trong cây, đại diện cho các thuộc tính được sử dụng để chia dữ liệu.
* **Nút lá (Leaf Nodes)**: Các nút cuối cùng, đại diện cho các nhãn hoặc giá trị dự đoán.

Quá trình xây dựng cây quyết định bao gồm việc chọn thuộc tính tốt nhất tại mỗi nút để chia dữ liệu, dựa trên các chỉ số như entropy và information gain.

### 2.1.2. Entropy và Information Gain

* **Entropy**: Entropy đo lường mức độ hỗn loạn hoặc không chắc chắn trong một tập hợp dữ liệu. Entropy cao khi dữ liệu không đồng nhất và thấp khi dữ liệu đồng nhất. Công thức tính entropy là:

Trong đó là xác suất của mỗi loại tình trạng sức khỏe.

* **Information Gain**: Information Gain đo lường mức giảm entropy khi một thuộc tính được sử dụng để chia dữ liệu. Công thức tính Information Gain là:

Trong đó:

* là entropy của tập dữ liệu trước khi chia.
* là kích thước của tập con .
* |D| là kích thước của toàn bộ tập dữ liệu.
* là entropy của tập con .

### 2.1.3. Mô hình học máy và trí tuệ nhân tạo

Học máy (Machine Learning) và trí tuệ nhân tạo (Artificial Intelligence) là các lĩnh vực khoa học máy tính nghiên cứu và phát triển các thuật toán và mô hình giúp máy tính có thể học và ra quyết định dựa trên dữ liệu. Các mô hình học máy có thể được chia thành hai nhóm chính: học máy giám sát (supervised learning) và học máy không giám sát (unsupervised learning).

Trong đề tài này, chúng ta sử dụng học máy giám sát để xây dựng mô hình cây quyết định, dựa trên các dữ liệu đầu vào (ví dụ: chiều cao, đường kính, tuổi của cây) và các nhãn tương ứng (tình trạng sức khỏe của cây).

## 2.2. Giả thuyết khoa học

Giả thuyết khoa học của đề tài này là việc áp dụng mô hình cây quyết định và các thuật toán học máy sẽ giúp cải thiện đáng kể hiệu quả quản lý cây xanh đô thị. Cụ thể, các mô hình này sẽ giúp dự đoán tình trạng sức khỏe của cây một cách chính xác hơn, từ đó giúp các cơ quan quản lý đưa ra các biện pháp chăm sóc và bảo dưỡng kịp thời, giảm thiểu các vấn đề sức khỏe của cây và đảm bảo an toàn cho người dân.

## 2.3. Phương pháp kỹ thuật và công nghệ

### 2.3.1. Ngôn ngữ lập trình và công nghệ web

* **Python**: Python là ngôn ngữ lập trình chính được sử dụng trong đề tài này do tính linh hoạt, dễ học và có nhiều thư viện hỗ trợ mạnh mẽ cho học máy và phân tích dữ liệu như Pandas, Scikit-learn, Plotly.
* **Streamlit**: Streamlit là một framework Python cho phép tạo ra các ứng dụng web tương tác một cách dễ dàng và nhanh chóng. Streamlit sẽ được sử dụng để phát triển giao diện người dùng cho hệ thống quản lý cây xanh.

### 2.3.2. Thư viện và công cụ hỗ trợ

* **Pandas**: Thư viện Pandas cung cấp các công cụ mạnh mẽ cho việc xử lý và phân tích dữ liệu. Pandas sẽ được sử dụng để thu thập, xử lý và làm sạch dữ liệu cây xanh.
* **Scikit-learn**: Scikit-learn là một thư viện học máy nổi tiếng trong Python, cung cấp nhiều thuật toán học máy, bao gồm cây quyết định. Thư viện này sẽ được sử dụng để xây dựng và huấn luyện mô hình cây quyết định.
* **Plotly**: Plotly là một thư viện Python cho phép tạo ra các biểu đồ tương tác. Plotly sẽ được sử dụng để tạo các biểu đồ phân tích và hiển thị thông tin cây xanh một cách trực quan.

### 2.3.3. Quy trình thực hiện đề tài

**Khảo sát và thu thập dữ liệu**: Thu thập dữ liệu từ các nguồn đáng tin cậy như cơ quan quản lý đô thị, các tổ chức môi trường và các cơ sở dữ liệu hiện có. Dữ liệu bao gồm các thông tin về các loài cây, đặc điểm vật lý của cây, tình trạng sức khỏe, và các sự kiện chăm sóc và bảo dưỡng.

**Xử lý và làm sạch dữ liệu**: Sử dụng thư viện Pandas để xử lý và làm sạch dữ liệu thu thập được. Quá trình này bao gồm việc kiểm tra tính đầy đủ và chính xác của dữ liệu, loại bỏ các giá trị thiếu hoặc không hợp lệ, và chuẩn hóa dữ liệu để phù hợp với các mô hình học máy.

**Phân tích và lựa chọn thuộc tính**: Sử dụng các phương pháp phân tích dữ liệu để xác định các thuộc tính quan trọng ảnh hưởng đến tình trạng sức khỏe của cây. Các thuộc tính này sẽ được sử dụng làm đầu vào cho mô hình cây quyết định.

**Xây dựng và huấn luyện mô hình cây quyết định**: Sử dụng thư viện Scikit-learn để xây dựng và huấn luyện mô hình cây quyết định. Quá trình huấn luyện bao gồm việc chia dữ liệu thành các tập huấn luyện và tập kiểm tra, điều chỉnh các tham số của mô hình để đạt được độ chính xác cao nhất.

**Kiểm thử và đánh giá mô hình**: Sử dụng dữ liệu kiểm tra để đánh giá hiệu suất của mô hình. Các chỉ số đánh giá bao gồm độ chính xác, độ nhạy, độ đặc hiệu và diện tích dưới đường cong ROC. Quá trình này giúp đảm bảo rằng mô hình có thể dự đoán tình trạng sức khỏe của cây một cách chính xác và tin cậy.

**Phát triển giao diện người dùng**: Sử dụng Streamlit để phát triển giao diện người dùng cho hệ thống quản lý cây xanh. Giao diện này sẽ cho phép người dùng nhập liệu, cập nhật thông tin cây xanh, xem các báo cáo và biểu đồ phân tích, và nhận các cảnh báo sớm về tình trạng sức khỏe của cây.

## 2.4. Đánh giá, nhận định tại sao lại chọn công nghệ/ngôn ngữ lập trình/framework đó để thực hiện đề tài

Trong việc xây dựng hệ thống quản lý cây xanh đô thị, việc lựa chọn công nghệ, ngôn ngữ lập trình và các framework hỗ trợ là một phần quan trọng để đảm bảo rằng hệ thống không chỉ hiệu quả mà còn dễ dàng phát triển và bảo trì. Các yếu tố như tính linh hoạt, hiệu suất, cộng đồng hỗ trợ và khả năng mở rộng đều được xem xét kỹ lưỡng. Dưới đây là các lý do chi tiết tại sao lại chọn các công nghệ và công cụ cụ thể như Python, Streamlit, Pandas, Scikit-learn và Plotly cho đề tài này.

**Python**

**Lý do chọn Python**:

* **Tính linh hoạt và dễ học**: Python là một ngôn ngữ lập trình bậc cao, dễ học và sử dụng, đặc biệt phù hợp cho cả những người mới bắt đầu và các chuyên gia lập trình. Cú pháp của Python rõ ràng, dễ đọc và dễ viết, giúp giảm thiểu lỗi lập trình và tăng năng suất làm việc.
* **Thư viện phong phú**: Python có một hệ sinh thái thư viện phong phú hỗ trợ nhiều lĩnh vực khác nhau, đặc biệt là học máy và phân tích dữ liệu. Các thư viện như Pandas, Scikit-learn, Matplotlib và Plotly cung cấp các công cụ mạnh mẽ để xử lý và phân tích dữ liệu, xây dựng và huấn luyện mô hình học máy, và tạo các biểu đồ trực quan.
* **Cộng đồng hỗ trợ lớn**: Python có một cộng đồng người dùng rộng lớn và năng động, cung cấp rất nhiều tài liệu, hướng dẫn và hỗ trợ trực tuyến. Điều này giúp dễ dàng tìm kiếm các giải pháp cho các vấn đề kỹ thuật và cập nhật với những phát triển mới nhất trong lĩnh vực.

**Tính năng nổi bật**:

* **Đa nền tảng**: Python có thể chạy trên nhiều hệ điều hành khác nhau như Windows, macOS và Linux, giúp hệ thống dễ dàng triển khai và sử dụng trên nhiều môi trường khác nhau.
* **Tương thích và tích hợp tốt**: Python dễ dàng tích hợp với các ngôn ngữ và công nghệ khác, cho phép phát triển các hệ thống phức tạp và đa nền tảng.

**Streamlit**

**Lý do chọn Streamlit**:

* **Dễ dàng phát triển giao diện người dùng**: Streamlit là một framework Python cho phép tạo ra các ứng dụng web tương tác một cách dễ dàng và nhanh chóng mà không cần kiến thức sâu về HTML, CSS hay JavaScript. Streamlit cung cấp các chức năng để tạo các form nhập liệu, hiển thị bảng dữ liệu, và tạo các biểu đồ tương tác một cách đơn giản và hiệu quả.
* **Tối ưu hóa cho phân tích dữ liệu và học máy**: Streamlit được thiết kế đặc biệt để hỗ trợ các dự án phân tích dữ liệu và học máy. Nó cho phép dễ dàng tích hợp với các thư viện học máy và phân tích dữ liệu như Pandas, Scikit-learn và Plotly, giúp nhanh chóng xây dựng và triển khai các mô hình học máy và trực quan hóa dữ liệu.

**Tính năng nổi bật**:

* **Thời gian thực**: Streamlit cung cấp khả năng cập nhật dữ liệu và mô hình học máy trong thời gian thực, giúp người dùng có thể thấy ngay các kết quả và phản hồi khi thay đổi các tham số hoặc dữ liệu đầu vào.
* **Tích hợp dễ dàng**: Streamlit dễ dàng tích hợp với các dịch vụ đám mây và các công cụ quản lý dự án, giúp đơn giản hóa quy trình triển khai và quản lý hệ thống.

**Pandas**

**Lý do chọn Pandas**:

* **Xử lý và phân tích dữ liệu mạnh mẽ**: Pandas cung cấp các cấu trúc dữ liệu và công cụ phân tích dữ liệu mạnh mẽ, cho phép dễ dàng thu thập, xử lý, làm sạch và chuyển đổi dữ liệu. Các chức năng của Pandas giúp xử lý dữ liệu nhanh chóng và hiệu quả, từ đó cải thiện năng suất làm việc.
* **Được sử dụng rộng rãi trong cộng đồng dữ liệu**: Pandas là một trong những thư viện phổ biến nhất trong cộng đồng khoa học dữ liệu và học máy, được sử dụng rộng rãi trong nhiều dự án phân tích dữ liệu và mô hình học máy.

**Tính năng nổi bật**:

* **Khả năng xử lý dữ liệu lớn**: Pandas có thể xử lý các tập dữ liệu lớn, giúp dễ dàng làm việc với các tập dữ liệu phức tạp và đa dạng.
* **Chức năng phong phú**: Pandas cung cấp nhiều chức năng tiện ích cho việc thao tác dữ liệu, từ việc đọc và ghi dữ liệu từ nhiều định dạng khác nhau, đến việc tính toán các thống kê mô tả, lọc dữ liệu, và thực hiện các phép biến đổi dữ liệu phức tạp.

**Scikit-learn**

**Lý do chọn Scikit-learn**:

* **Thư viện học máy toàn diện**: Scikit-learn cung cấp một tập hợp phong phú các thuật toán học máy, từ các thuật toán cơ bản đến các thuật toán nâng cao, giúp dễ dàng xây dựng và huấn luyện các mô hình học máy. Thư viện này hỗ trợ nhiều thuật toán phân loại, hồi quy, clustering, và giảm chiều dữ liệu.
* **Tài liệu và hỗ trợ tốt**: Scikit-learn có tài liệu chi tiết và dễ hiểu, cung cấp nhiều ví dụ và hướng dẫn sử dụng, giúp người dùng nhanh chóng làm quen và áp dụng vào các dự án của mình.

**Tính năng nổi bật**:

* **Tích hợp tốt với các thư viện khác**: Scikit-learn dễ dàng tích hợp với các thư viện khác trong hệ sinh thái Python như Pandas, NumPy và Matplotlib, giúp xây dựng các quy trình học máy liền mạch và hiệu quả.
* **Hiệu suất cao**: Scikit-learn được tối ưu hóa để xử lý các tập dữ liệu lớn và phức tạp, giúp cải thiện hiệu suất và độ chính xác của các mô hình học máy.

**Plotly**

**Lý do chọn Plotly**:

* **Tạo biểu đồ tương tác**: Plotly cho phép tạo ra các biểu đồ tương tác với khả năng tùy chỉnh cao, giúp trực quan hóa dữ liệu một cách sinh động và dễ hiểu. Các biểu đồ của Plotly có thể tích hợp trực tiếp vào các ứng dụng web, giúp người dùng dễ dàng tương tác và khám phá dữ liệu.
* **Hỗ trợ nhiều loại biểu đồ**: Plotly hỗ trợ nhiều loại biểu đồ khác nhau như biểu đồ cột, biểu đồ đường, biểu đồ scatter, biểu đồ bánh, và nhiều loại biểu đồ phức tạp khác, đáp ứng đa dạng nhu cầu trực quan hóa dữ liệu.

**Tính năng nổi bật**:

* **Tương thích với các công cụ phân tích dữ liệu**: Plotly dễ dàng tích hợp với các công cụ phân tích dữ liệu như Pandas, NumPy và Scikit-learn, giúp tạo ra các biểu đồ phân tích từ các mô hình học máy và dữ liệu đầu vào một cách hiệu quả.
* **Khả năng tùy chỉnh cao**: Plotly cung cấp nhiều tùy chọn tùy chỉnh biểu đồ, từ việc thay đổi màu sắc, kiểu đường, nhãn đến việc thêm các chú thích và các yếu tố tương tác, giúp tạo ra các biểu đồ trực quan và chuyên nghiệp.

# Chương 3: Giải quyết bài toán

## 3.1. Phân tích các yêu cầu

Bước đầu tiên trong việc giải quyết bài toán là phân tích các yêu cầu cụ thể của hệ thống quản lý cây xanh đô thị. Điều này bao gồm xác định rõ các yêu cầu về dữ liệu, chức năng xử lý và quy trình xử lý.

### 3.1.1. Yêu cầu về dữ liệu

* **Thông tin cây xanh**: Bao gồm các đặc điểm vật lý như chiều cao, đường kính, tuổi cây, loài cây, tình trạng sức khỏe, vị trí, ghi chú và ngày nhắc nhở.
* **Thông tin về loài cây**: Cung cấp các thông tin cơ bản về từng loài cây như tên loài, đặc điểm sinh học, điều kiện sống, và các thông tin liên quan.

### 3.1.2. Yêu cầu về chức năng xử lý

* **Theo dõi và cập nhật thông tin cây xanh**: Hệ thống cần cho phép người dùng cập nhật thông tin về cây xanh, bao gồm các đặc điểm vật lý và tình trạng sức khỏe hiện tại của cây.
* **Dự đoán tình trạng sức khỏe của cây**: Sử dụng các thuật toán học máy để dự đoán tình trạng sức khỏe của cây dựa trên các đặc điểm vật lý và dữ liệu lịch sử.
* **Tạo báo cáo và biểu đồ phân tích**: Hệ thống cần cung cấp các báo cáo, biểu đồ trực quan giúp người dùng dễ dàng theo dõi và phân tích tình trạng sức khỏe của cây xanh.

### 3.1.3. Yêu cầu về quy trình xử lý

* **Thu thập dữ liệu**: Dữ liệu cần được thu thập từ các nguồn đáng tin cậy như cơ quan quản lý đô thị, các tổ chức môi trường và các cơ sở dữ liệu hiện có.
* **Xử lý dữ liệu**: Dữ liệu cần được xử lý và làm sạch trước khi sử dụng trong các thuật toán học máy.
* **Huấn luyện mô hình**: Sử dụng các thuật toán học máy để huấn luyện mô hình dự đoán tình trạng sức khỏe của cây.
* **Kiểm thử và đánh giá**: Hệ thống cần được kiểm thử và đánh giá trên các dữ liệu thực tế để đảm bảo độ chính xác và hiệu quả.

## 3.2. Thiết kế hệ thống

Thiết kế hệ thống bao gồm việc xác định kiến trúc hệ thống, thiết kế cơ sở dữ liệu và thiết kế giao diện người dùng.

### 3.2.1. Kiến trúc hệ thống

Hệ thống được thiết kế theo kiến trúc phân lớp, bao gồm các lớp chính như sau:

* **Lớp giao diện người dùng (UI Layer)**: Chịu trách nhiệm hiển thị và tương tác với người dùng. Lớp này sẽ được phát triển bằng Streamlit, cho phép tạo ra các ứng dụng web tương tác một cách dễ dàng.
* **Lớp logic ứng dụng (Application Logic Layer)**: Chịu trách nhiệm xử lý các yêu cầu từ người dùng, thực hiện các thuật toán học máy và quản lý luồng dữ liệu. Lớp này sẽ được phát triển bằng Python và sử dụng các thư viện như Pandas và Scikit-learn.
* **Lớp dữ liệu (Data Layer)**: Chịu trách nhiệm lưu trữ và quản lý dữ liệu. Dữ liệu sẽ được lưu trữ trong cơ sở dữ liệu quan hệ và được truy vấn thông qua các công cụ hỗ trợ như SQLAlchemy.

### 3.2.2. Thiết kế cơ sở dữ liệu

Cơ sở dữ liệu của hệ thống sẽ bao gồm các bảng chính như sau:

* **Bảng Tree**: Lưu trữ thông tin về cây, bao gồm các cột như TreeId, SpeciesId, Age, Height, Diameter, HealthStatus, Note, Location, ReminderDate.
* **Bảng Species**: Lưu trữ thông tin về loài cây, bao gồm các cột như SpeciesId, SpeciesName.

### **3.2.3. Thiết kế giao diện người dùng**

Giao diện người dùng (UI) được thiết kế để đảm bảo tính thân thiện và dễ sử dụng, giúp người dùng dễ dàng truy cập và thao tác với hệ thống quản lý cây xanh. Giao diện sẽ bao gồm các trang chính như sau:

* **Trang danh sách cây xanh**:
  + **Hiển thị danh sách cây xanh**: Giao diện sẽ hiển thị một bảng danh sách các cây xanh, bao gồm các thông tin như tên loài cây, chiều cao, đường kính, tuổi, tình trạng sức khỏe, vị trí, và ghi chú.
  + **Chức năng tìm kiếm và lọc**: Người dùng có thể tìm kiếm cây xanh dựa trên các tiêu chí như tên loài cây, tình trạng sức khỏe hoặc vị trí. Các bộ lọc này giúp người dùng nhanh chóng tìm thấy thông tin cần thiết.
  + **Xem chi tiết cây xanh**: Khi người dùng chọn một cây từ danh sách, hệ thống sẽ hiển thị chi tiết thông tin của cây đó, bao gồm các đặc điểm vật lý và lịch sử chăm sóc.
* **Trang thêm cây mới**:
  + **Nhập liệu thông tin cây mới**: Người dùng có thể nhập các thông tin về cây mới như tên loài cây, chiều cao, đường kính, tuổi, tình trạng sức khỏe, vị trí, và ghi chú. Giao diện sẽ cung cấp các ô nhập liệu tương ứng cho từng thông tin.
  + **Xác nhận và lưu thông tin**: Sau khi nhập liệu, người dùng có thể xác nhận để lưu thông tin cây mới vào cơ sở dữ liệu. Hệ thống sẽ kiểm tra tính hợp lệ của dữ liệu trước khi lưu.
* **Trang chỉnh sửa và xóa cây**:
  + **Chỉnh sửa thông tin cây**: Người dùng có thể chọn một cây từ danh sách để chỉnh sửa thông tin. Giao diện chỉnh sửa sẽ hiển thị các ô nhập liệu tương ứng với thông tin hiện tại của cây, cho phép người dùng thay đổi và cập nhật thông tin.
  + **Xóa cây**: Người dùng có thể chọn một cây từ danh sách và yêu cầu xóa cây. Hệ thống sẽ yêu cầu xác nhận trước khi xóa cây khỏi cơ sở dữ liệu để đảm bảo không có sai sót.
* **Trang báo cáo và biểu đồ**:
  + **Báo cáo tình trạng cây xanh**: Giao diện sẽ hiển thị các báo cáo tổng quan về tình trạng cây xanh trong khu vực, bao gồm số lượng cây theo tình trạng sức khỏe, tuổi cây, và loại cây.
  + **Biểu đồ phân tích**: Giao diện sẽ hiển thị các biểu đồ phân tích như biểu đồ số lượng cây theo tình trạng sức khỏe, biểu đồ phân bố chiều cao và đường kính cây, và biểu đồ tỷ lệ loài cây. Các biểu đồ này giúp người dùng dễ dàng nắm bắt thông tin và đưa ra các quyết định quản lý.

## **3.3. Cài đặt hệ thống**

Quá trình cài đặt hệ thống bao gồm các bước chính như thiết lập môi trường phát triển, cài đặt các thư viện và công cụ hỗ trợ, và triển khai các thành phần của hệ thống.

### **3.3.1. Thiết lập môi trường phát triển**

* **Cài đặt Python**: Cài đặt Python trên máy phát triển. Đảm bảo rằng phiên bản Python sử dụng tương thích với các thư viện và công cụ hỗ trợ.

### **3.3.2. Cài đặt các thư viện và công cụ hỗ trợ**

* **Pandas:** Cài đặt thư viện Pandas để xử lý và phân tích dữ liệu.
* **Scikit-learn**: Cài đặt thư viện Scikit-learn để xây dựng và huấn luyện mô hình học máy.
* **Plotly**: Cài đặt thư viện Plotly để tạo các biểu đồ tương tác.
* **Streamlit**: Cài đặt Streamlit để phát triển giao diện người dùng.

### **3.3.3. Triển khai các thành phần của hệ thống**

* **Xây dựng mô hình cây quyết định**: Sử dụng Scikit-learn để xây dựng và huấn luyện mô hình cây quyết định dựa trên dữ liệu cây xanh. Lưu mô hình sau khi huấn luyện để sử dụng trong dự đoán.
* **Phát triển giao diện người dùng**: Sử dụng Streamlit để phát triển các trang giao diện như danh sách cây xanh, thêm cây mới, chỉnh sửa và xóa cây, và báo cáo biểu đồ. Tích hợp các chức năng xử lý dữ liệu và mô hình học máy vào giao diện người dùng.
* **Kết nối cơ sở dữ liệu**: Sử dụng SQLAlchemy để kết nối và tương tác với cơ sở dữ liệu. Đảm bảo rằng các thao tác truy vấn, thêm mới, cập nhật, và xóa dữ liệu đều hoạt động chính xác.

# **Chương 4: Kết quả đạt được**

Chương này trình bày chi tiết các kết quả thực nghiệm từ hệ thống quản lý cây xanh đô thị, bao gồm các phân tích, dự đoán và đánh giá tình trạng sức khỏe cây xanh, cùng với các chú thích và đánh giá về hiệu quả của hệ thống.

## **4.1. Thu thập và xử lý dữ liệu**

Quá trình thu thập dữ liệu được thực hiện từ nhiều nguồn đáng tin cậy bao gồm các cơ quan quản lý đô thị, các tổ chức môi trường và các cơ sở dữ liệu hiện có. Dữ liệu bao gồm thông tin chi tiết về các loài cây, các đặc điểm vật lý của cây (chiều cao, đường kính, tuổi), và tình trạng sức khỏe của cây. Dữ liệu sau khi thu thập được xử lý và làm sạch để đảm bảo tính chính xác và nhất quán.

**1. Dữ liệu cây xanh:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Loài cây** | **Chiều cao (m)** | **Đường kính (cm)** | **Tuổi (năm)** | **Tình trạng sức khỏe** |
| Sồi | 10 | 30 | 5 | Khỏe mạnh |
| Bạch Đàn | 15 | 50 | 10 | Yếu |
| Tre | 20 | 60 | 15 | Khỏe mạnh |
| Keo | 25 | 80 | 20 | Bệnh |
| Hoa Sữa | 30 | 100 | 25 | Cần tỉa |

**2. Các tính toán và phân tích chi tiết:**

**2.1. Tính chiều cao trung bình theo tình trạng sức khỏe**

**Công thức:** Chiều cao trung bình = (Tổng chiều cao của cây thuộc tình trạng sức khỏe) / (Số lượng cây trong tình trạng sức khỏe).

|  |  |  |  |
| --- | --- | --- | --- |
| Tình trạng sức khỏe | Tổng chiều cao (m) | Số lượng cây | Chiều cao trung bình (m) |
| Khỏe mạnh | 10 + 20 | 2 | 15 |
| Yếu | 15 | 1 | 15 |
| Bệnh | 25 | 1 | 25 |
| Cần tỉa | 30 | 1 | 30 |

**2.2. Tính tuổi trung bình theo tình trạng sức khỏe**

**Công thức:** Tuổi trung bình = (Tổng tuổi của cây trong từng tình trạng sức khỏe) / (Số lượng cây trong tình trạng sức khỏe).

|  |  |  |  |
| --- | --- | --- | --- |
| Tình trạng sức khỏe | Tổng tuổi (năm) | Số lượng cây | Tuổi trung bình (năm) |
| Khỏe mạnh | 5 + 15 | 2 | 10 |
| Yếu | 10 | 1 | 10 |
| Bệnh | 20 | 1 | 20 |
| Cần tỉa | 25 | 1 | 25 |

**2.3. Tính đường kính trung bình theo tình trạng sức khỏe**

**Công thức:** Đường kính trung bình = (Tổng đường kính của cây thuộc tình trạng sức khỏe) / (Số lượng cây trong tình trạng sức khỏe).

|  |  |  |  |
| --- | --- | --- | --- |
| Tình trạng sức khỏe | Tổng đường kính (cm) | Số lượng cây | Đường kính trung bình (cm) |
| Khỏe mạnh | 30 + 60 | 2 | 45 |
| Yếu | 50 | 1 | 50 |
| Bệnh | 80 | 1 | 80 |
| Cần tỉa | 100 | 1 | 100 |

## **4.2. Huấn luyện mô hình cây quyết định**

Sử dụng mô hình cây quyết định tự định nghĩa dựa trên thư viện Scikit-learn, mô hình được huấn luyện dựa trên tập dữ liệu đã được xử lý. Quá trình huấn luyện bao gồm việc chia dữ liệu thành các tập huấn luyện và kiểm tra, điều chỉnh các tham số của mô hình để đạt được độ chính xác cao nhất.

A screen shot of a computer code

Description automatically generated

Hình 4.1 Lớp DecisionTreeClassifierCustom

* Lớp DecisionTreeClassifierCustom

Được sử dụng để tạo và huấn luyện mô hình cây quyết định.

self.tree khởi tạo cây quyết định dưới dạng một từ điển.

A computer screen shot of a code

Description automatically generated

Hình 4.2 Phương thức fit

* Phương thức fit:

Dùng để huấn luyện mô hình cây quyết định với dữ liệu đầu vào x (các đặc điểm cây) và y (nhãn tình trạng sức khỏe).

Sao chép x và thêm y vào dưới dạng cột label.

Gọi phương thức \_build\_tree để xây dựng cây quyết định.

A computer code with text

Description automatically generated with medium confidence

Hình 4.3 Phương thức \_entropy

* Phương thức \_entropy:

Tính toán entropy của tập dữ liệu y.

counts là số lượng xuất hiện của từng nhãn trong y.

probabilities là tỉ lệ xuất hiện của từng nhãn.

A computer screen shot of a code

Description automatically generated

Hình 4.4 Phương thức \_information\_gain

* Phương thức \_information\_gain:

Tính toán độ lợi thông tin (information gain) cho thuộc tính feature trong tập dữ liệu data.

values là các giá trị duy nhất của thuộc tính feature.

Tính toán entropy có trọng số cho từng giá trị của feature.

Độ lợi thông tin là sự chênh lệch giữa entropy của nhãn và entropy có trọng số.

A screen shot of a computer code

Description automatically generated

Hình 4.5 Phương thức \_best\_feature

* Phương thức \_best\_feature:

Tìm thuộc tính tốt nhất để chia dữ liệu dựa trên độ lợi thông tin.

Tính toán entropy của nhãn (label\_entropy).

Tính toán độ lợi thông tin cho từng thuộc tính.

Trả về thuộc tính có độ lợi thông tin cao nhất.

A computer screen shot of a code

Description automatically generated

Hình 4.6 Phương thức \_build\_tree

* Phương thức \_build\_tree:

Xây dựng cây quyết định đệ quy từ tập dữ liệu data.

Kiểm tra nếu tất cả các nhãn trong data đều giống nhau, trả về nhãn đó.

Nếu chỉ còn một thuộc tính, trả về nhãn xuất hiện nhiều nhất.

Tìm thuộc tính tốt nhất để chia dữ liệu (best\_feature).

Tạo cây con cho mỗi giá trị của thuộc tính tốt nhất và gọi đệ quy \_build\_tree.

A computer screen shot of text

Description automatically generated

Hình 4.7 Phương thức predict và \_predict\_row

* Phương thức predict:

Dự đoán nhãn cho tập dữ liệu mới x.

Áp dụng \_predict\_row cho từng hàng trong x.

* Phương thức \_predict\_row:

Dự đoán nhãn cho một hàng dữ liệu row dựa trên cây quyết định tree.

Nếu cây đã đến nút lá, trả về giá trị của nút lá.

Nếu không, di chuyển đến nhánh tương ứng với giá trị của thuộc tính hiện tại và tiếp tục dự đoán.

A computer screen shot of a program code

Description automatically generated

Hình 4.8 Phương thức train\_decision\_tree

* Phương thức train\_decision\_tree:

Chuyển đổi dữ liệu đầu vào thành một DataFrame của Pandas để dễ dàng thao tác và xử lý dữ liệu.

Kiểm tra xem DataFrame có rỗng hay không. Nếu không rỗng, tiếp tục thực hiện các bước tiếp theo.

Lấy các cột Height (Chiều cao), Diameter (Đường kính) và Age (Tuổi) từ DataFrame và lưu vào biến x. Đây là các đặc điểm của cây sẽ được sử dụng làm biến đầu vào cho mô hình.

Lấy cột HealthStatus (Tình trạng sức khỏe) từ DataFrame và lưu vào biến y. Đây là biến mục tiêu mà mô hình sẽ dự đoán.

Sử dụng np.unique để mã hóa các giá trị tình trạng sức khỏe thành nhãn số. y\_labels chứa các nhãn nguyên gốc, còn y chứa các nhãn số tương ứng với từng tình trạng sức khỏe.

Tạo và huấn luyện mô hình cây quyết định tùy chỉnh với dữ liệu đầu vào x và các nhãn số y. Trả về mô hình cây quyết định đã huấn luyện (clf) và các nhãn nguyên gốc (y\_labels).

## **4.3. Phân tích và dự đoán tình trạng sức khỏe cây**

Sử dụng mô hình cây quyết định tự định nghĩa dựa theo thư viện Scikit-learn, hệ thống thực hiện phân tích và dự đoán tình trạng sức khỏe của các cây xanh trong tập dữ liệu. Kết quả phân tích bao gồm các báo cáo và biểu đồ trực quan, giúp người dùng dễ dàng theo dõi và đánh giá tình trạng sức khỏe của cây xanh.

**Biểu đồ phân bố chiều cao và đường kính cây**:

Biểu đồ phân bố chiều cao và đường kính cây cho thấy sự phân bố đồng đều của các đặc điểm vật lý của cây trong tập dữ liệu.

**Biểu đồ tỷ lệ loài cây**:

Biểu đồ tỷ lệ loài cây cung cấp cái nhìn tổng quan về sự đa dạng và phân bố các loài cây trong khu vực quản lý.

## **4.4. Đánh giá kết quả đạt được**

Quá trình kiểm thử và đánh giá hệ thống cho thấy các kết quả đạt được có tính khả thi và chính xác cao. Hệ thống đã thể hiện khả năng dự đoán tình trạng sức khỏe của cây một cách hiệu quả, cung cấp các báo cáo và biểu đồ trực quan giúp người dùng dễ dàng nắm bắt thông tin và đưa ra các quyết định quản lý.

**Ưu điểm**:

* **Độ chính xác cao**: Mô hình cây quyết định tự định nghĩa đạt độ chính xác cao trong việc dự đoán tình trạng sức khỏe của cây, giúp giảm thiểu các sai sót và đưa ra các biện pháp xử lý kịp thời.
* **Giao diện thân thiện**: Giao diện người dùng được thiết kế thân thiện, dễ sử dụng, giúp người dùng dễ dàng thao tác và truy cập thông tin.
* **Cảnh báo sớm**: Hệ thống có chức năng cảnh báo sớm về các vấn đề sức khỏe của cây, giúp người quản lý có biện pháp xử lý kịp thời, giảm thiểu rủi ro và đảm bảo an toàn.

**Nhược điểm**:

* **Phụ thuộc vào dữ liệu**: Độ chính xác của mô hình phụ thuộc vào chất lượng và độ chính xác của dữ liệu đầu vào. Việc thu thập và xử lý dữ liệu cần được thực hiện cẩn thận để đảm bảo tính chính xác và nhất quán.
* **Giới hạn trong phạm vi nghiên cứu**: Hiện tại, hệ thống mới được kiểm thử trên một phạm vi nghiên cứu cụ thể. Để áp dụng rộng rãi, hệ thống cần được kiểm thử và điều chỉnh trên nhiều khu vực khác nhau.

# Kết luận

**Kết quả đạt được**

Đồ án đã đạt được nhiều kết quả quan trọng, đóng góp tích cực vào việc quản lý và bảo vệ cây xanh đô thị. Dưới đây là những kết quả chính đã đạt được:

1. **Xây dựng hệ thống quản lý cây xanh đô thị**:
   * Hệ thống đã được phát triển và triển khai thành công, bao gồm các chức năng chính như theo dõi, cập nhật thông tin cây xanh, dự đoán tình trạng sức khỏe cây, và tạo báo cáo phân tích.
   * Giao diện người dùng thân thiện, dễ sử dụng, giúp người dùng dễ dàng thao tác và truy cập thông tin cây xanh.
2. **Huấn luyện và áp dụng mô hình cây quyết định tự định nghĩa**:
   * Mô hình cây quyết định đã được xây dựng và huấn luyện thành công, đạt được độ chính xác cao trong việc dự đoán tình trạng sức khỏe của cây.
3. **Phân tích và dự đoán tình trạng sức khỏe cây xanh**:
   * Hệ thống đã thực hiện phân tích và dự đoán tình trạng sức khỏe của các cây xanh trong tập dữ liệu, cung cấp các báo cáo và biểu đồ trực quan giúp người dùng dễ dàng theo dõi và đánh giá tình trạng sức khỏe của cây.

**Kiến nghị về những nghiên cứu, phát triển tiếp theo**

Mặc dù đã đạt được nhiều kết quả tích cực, nhưng hệ thống quản lý cây xanh đô thị vẫn còn nhiều tiềm năng để phát triển và cải thiện. Dưới đây là một số kiến nghị về các nghiên cứu và phát triển tiếp theo:

1. **Cải thiện quá trình thu thập dữ liệu**:
   * Để đảm bảo tính chính xác và đầy đủ của dữ liệu, cần thiết lập các quy trình thu thập dữ liệu chặt chẽ hơn và sử dụng các công cụ hỗ trợ tự động hóa. Việc áp dụng công nghệ IoT (Internet of Things) có thể giúp thu thập dữ liệu theo thời gian thực, nâng cao độ chính xác và kịp thời của dữ liệu.
2. **Nâng cao độ chính xác của mô hình học máy**:
   * Tiếp tục nghiên cứu và phát triển các thuật toán học máy mới nhằm cải thiện độ chính xác và hiệu quả của mô hình dự đoán. Các kỹ thuật học sâu (deep learning) và mạng nơ-ron nhân tạo (artificial neural networks) có thể được xem xét để áp dụng vào hệ thống.
3. **Mở rộng phạm vi nghiên cứu và triển khai hệ thống**:
   * Kiểm thử và triển khai hệ thống trên nhiều khu vực khác nhau để đảm bảo tính khả thi và hiệu quả của hệ thống trong các điều kiện thực tế đa dạng. Việc áp dụng hệ thống vào các thành phố lớn và khu vực có điều kiện môi trường khác nhau sẽ giúp đánh giá và cải thiện tính năng của hệ thống.
4. **Phát triển các chức năng mới**:
   * Phát triển thêm các chức năng mới như quản lý sự kiện liên quan đến cây xanh (chặt hạ, trồng mới), tích hợp với hệ thống bản đồ để hiển thị vị trí cây xanh, và cung cấp các thông tin khoa học về loài cây và điều kiện sống.
5. **Nâng cao khả năng tương tác và hỗ trợ người dùng**:
   * Phát triển các tính năng tương tác cao hơn như chatbot hỗ trợ người dùng, cung cấp các hướng dẫn chăm sóc cây xanh tự động, và tạo diễn đàn cho người dùng trao đổi và chia sẻ kinh nghiệm.
6. **Đánh giá và cải thiện hệ thống dựa trên phản hồi người dùng**:
   * Thu thập phản hồi từ người dùng để đánh giá hiệu quả và tính năng của hệ thống. Dựa trên các phản hồi này, tiếp tục cải thiện và nâng cao hệ thống để đáp ứng tốt hơn nhu cầu của người dùng.