TRƯỜNG ĐẠI HỌC HẠ LONG
KHOA CÔNG NGHỆ THÔNG TIN

- - - - - - - -

ĐẠI HỌC HẠ LONG
Học để thành công

BÁO CÁO KẾT THÚC HỌC PHẦN

TÊN ĐỀ TÀI:

TÌM HIỂU VÀ XÂY DỰNG MÔ HÌNH DECISION TREE CHO DỰ ĐOÁN LOẠI LÚA MÌ

Quảng Ninh, tháng 11 năm 2023

TRƯỜNG ĐẠI HỌC HẠ LONG
KHOA CÔNG NGHỆ THÔNG TIN

- - - - ❀ - - - -

BÁO CÁO KẾT THÚC HỌC PHẦN

TÊN ĐỀ TÀI:

TÌM HIỂU VÀ XÂY DỰNG MÔ HÌNH DECISION TREE CHO DỰ ĐOÁN LOẠI LÚA MÌ

Sinh viên thực hiện: **Nguyễn Văn Thạch**
Phạm Minh Thiên
Toly Keopaserth

Giảng viên hướng dẫn: **TS. Nguyễn Văn Hậu**

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành nhất đến Thầy Nguyễn Văn Hậu đã hỗ trợ và đồng hành cùng chúng em trong suốt khoá học máy này. Đây là một hành trình học tập đầy thách thức và hứa hẹn, và chúng em rất biết ơn vì sự hướng dẫn và giáo dục tận tâm mà chúng em đã nhận được.

Trải qua bao nhiêu tuần học tập, chúng em đã có cơ hội khám phá và áp dụng những kiến thức lý thuyết học máy vào thực tế. Các bài giảng, bài thực hành và dự án đã giúp chúng em hiểu rõ hơn về cách xây dựng và triển khai mô hình máy học. Đặc biệt, những phản hồi chi tiết từ Thầy Nguyễn Văn Hậu đã là nguồn động viên quý báu, giúp chúng em khắc phục những khó khăn và phát triển kỹ năng một cách liên tục.

Chúng em cũng muốn bày tỏ lòng biết ơn sâu sắc đến sự hỗ trợ không ngừng từ bạn bè và đồng học. Sự chia sẻ kiến thức và kinh nghiệm giữa các thành viên trong nhóm đã tạo nên một môi trường học tập tích cực và sáng tạo.

Cuối cùng, chúng em xin cam kết sẽ tiếp tục áp dụng những kiến thức và kỹ năng học được từ khoá học này vào công việc thực tế. Báo cáo cuối học phần không chỉ là kết quả của sự nỗ lực cá nhân mà còn là sự góp phần vào sự thành công của cả nhóm.

Một lần nữa, chúng em xin chân thành cảm ơn Thầy Nguyễn Văn Hậu và mọi người đã giúp chúng em trở thành những người học máy tự tin và có khả năng đối mặt với thách thức của thế giới công nghiệp 4.0 ngày nay.

Sinh viên	Sinh viên	Sinh viên
Phạm Minh Thiên	Nguyễn Văn Thạch	Toly Keopaserth

MỤC LỤC

MỤC LỤC 3
CHƯƠNG 1. TỔNG QUAN BÀI TOÁN 6
1.1. Tổng quan về Máy học 6
1.1.1. Máy học là gì? 6
1.1.2. Các phương pháp Machine Learning 6
1.1.3. Phân loại các thuật toán Machine Learning 7
1.1.4. Các ứng dụng của Machine Learning 9
1.2. Bài toán dự đoán loại lúa mì 9
1.2.1. Giới thiệu bài toán 9
1.2.2. Dữ liệu 10
1.2.3. Mục tiêu 10
1.2.4. Các bước thực hiện 10
1.2.5. Ứng dụng thực tế 10
CHƯƠNG 2. MÔ HÌNH HỌC MÁY 12
2.1. Ý tưởng mô hình 12
2.2. Phân tích mô hình 12
2.2.1. Biểu diễn cây quyết định 12
2.2.2. Quyết định tại các nút 13
2.2.3. Hiệu suất mô hình 13
2.2.4. Diễn giải quyết định 13
2.2.5. Visualize cây 13
2.2.6. Xử lý Overfitting và Underfitting 13
2.2.7. Ưu, nhược điểm của mô hình 14
2.2.8. Ví dụ 14
2.3. Cách thức đánh giá mô hình 16
2.3.1. Chia dữ liệu 16
2.3.2. Huấn luyện mô hình 16
2.3.3. Kiểm thử trên tập kiểm thử 16
2.3.4. Đánh giá hiệu suất 16

CHƯƠNG 3. CÀI ĐẶT MÔ HÌNH 17
3.1. Dữ liệu bài toán 17
3.2. Hiển thị dữ liệu 21
3.2.1. Mã chương trình 21
3.2.2. Kết quả 21
3.3. Cài đặt mô hình cây quyết định bằng Python (sklearn) 21
3.3.1 Mã chương trình 22
3.3.2 Kết quả 25
3.3.3 Các biểu đồ 26
3.4. Đánh giá độ chính xác của mô hình 31
3.4.1 Mã chương trình 31
3.4.2 Kết quả 31
3.4.3 Kết luận 32
TỔNG KẾT 33
1. Các nhiệm vụ đã thực hiện 33
2. Ưu điểm và nhược điểm của chương trình 33
TÀI LIỆU THAM KHẢO 35

CHƯƠNG 1. TỔNG QUAN BÀI TOÁN

1.1. Tổng quan về Máy học

1.1.1. Máy học là gì?

Học máy – Machine learning (ML) là tập hợp con của trí tuệ nhân tạo (AI) cung cấp cho máy móc hoặc chương trình máy tính học hỏi từ kinh nghiệm (những gì đã được học) của chính chúng hoặc dữ liệu có sẵn mà không cần lập trình cụ thể.

Nói cách khác, Machine Learning là quá trình đào tạo máy tính bắt chước hành vi của con người. Trong khi con người học hỏi từ kinh nghiệm và thông qua các giác quan thì máy móc hoặc máy tính học hỏi từ dữ liệu.

Càng cung cấp nhiều dữ liệu cho giải pháp Machine Learning thì kết quả nhận về càng tốt và chính xác hơn. Machine Learning phát triển một quy trình suy nghĩ tự chủ theo thời gian, cho phép nó hoàn thành các nhiệm vụ và quy trình kinh doanh khác nhau mà không cần giám sát (nhưng một số mô hình yêu cầu giám sát).

1.1.2. Các phương pháp Machine Learning

Supervised Learning

Supervised Learning là một kỹ thuật học máy được sử dụng trong sản xuất để đào tạo các mô hình sử dụng dữ liệu được dán nhãn. Trong sản xuất, Supervised Learning có thể được sử dụng để phân loại lỗi, xác định các thông số sản xuất và dự đoán thời gian sử dụng hữu ích còn lại của thiết bị.

Unsupervised Learning

Unsupervised Learning là một kỹ thuật học máy được sử dụng trong sản xuất để đào tạo các mô hình sử dụng dữ liệu chưa được gắn nhãn. Phương pháp này dựa vào dữ liệu được dán nhãn và các kỹ thuật phân cụm để xác định các mẫu và cấu trúc trong dữ liệu. Trong sản xuất, Unsupervised Learning có thể được sử dụng để xác

định các nhóm sản phẩm tương tự hoặc xác định các khiếm khuyết khó phát hiện.

Deep learning

Deep Learning là một tập hợp con của học máy sử dụng mạng lưới thần kinh để mô hình hóa các mẫu phức tạp trong dữ liệu. Trong sản xuất, Deep Learning có thể được sử dụng để mô hình hóa các mối quan hệ phức tạp giữa các thông số sản xuất và chất lượng sản phẩm. Nó cũng có thể được sử dụng để phân tích lượng lớn dữ liệu từ cảm biến và các nguồn khác nhằm dự đoán lỗi thiết bị.

Reinforcement Learning

Reinforcement Learning là một kỹ thuật học máy được sử dụng trong sản xuất để dạy các mô hình đưa ra quyết định dựa trên phản hồi từ môi trường. Trong sản xuất, Reinforcement Learning có thể được sử dụng để tối ưu hóa quy trình sản xuất và cải thiện việc bảo trì thiết bị.

1.1.3. Phân loại các thuật toán Machine Learning

Dựa trên các tiêu chí khác nhau, người ta có thể phân loại các thuật toán Học máy theo nhiều cách khác nhau. Chẳng hạn, dựa vào vấn đề, nhiệm vụ cần giải quyết của thuật toán, người ta phân loại các thuật toán Học máy thành ba loại:

- Hồi quy (Regression): Giải quyết bài toán dự đoán giá trị một đại lượng nào đó dựa vào giá trị của các đại lượng liên quan. Ví dụ, dựa vào các đặc điểm như diện tích, số phòng, khoảng cách tới trung tâm…để dự đoán giá trị căn nhà.
- Phân lớp (Classification): Giải quyết các bài toán nhận dạng xem một đối tượng thuộc lớp nào trong số các lớp cho trước. Ví dụ, bài toán nhận diện chữ viết, bài toán phân loại email…thuộc các thuật toán phân lớp.

- Phân cụm (Clustering): Ý tưởng cơ bản giống với các thuật toán phân lớp, sự khác biệt là ở chỗ, trong các bài toán phân cụm, các cụm chưa được xác định trước và thuật toán phải tự khám phá và phân cụm dữ liệu.

Dựa trên cách máy tính học, người ta chia các thuật toán Học máy thành:

- Học có giám sát (Supervised learning): Thuật toán sẽ học trên dữ liệu đã được dán nhãn. Ví dụ, trong bài toán nhận diện hình ảnh, dữ liệu đầu vào sẽ là rất nhiều bức ảnh khác nhau về loài mèo. Thuật toán sẽ học các đặc điểm quan trọng từ các bức ảnh đó để nhận biết xem một đối tượng trong một bức ảnh có phải là mèo hay không.
- Học không giám sát (Unsupervised learning): Thuật toán học trên các dữ liệu chưa được gán nhãn và sẽ phải tự khám phá ra cấu trúc, phân bố của dữ liệu để tự phân cụm chúng.
- Học bán giám sát (Semi-supervised learning): Kết hợp cả học giám sát và học không giám sát. Tức là, một số dữ liệu đầu vào sẽ được gán nhãn và một số khác thì không được gán nhãn.
- Học tăng cường/củng cố (Reinforced learning): Thuật toán sẽ tự học dựa trên việc tính điểm thưởng, phạt cho các kết quả thực hiện nhiệm vụ. Cụ thể hơn, các thuật toán học tăng cường nghiên cứu cách thức một tác nhân (Agent) trong một môi trường (Environment) đang ở một trạng thái (State) thực hiện một hành động (Action) để tối ưu hóa một phần thưởng (Reward) chung. Các chương trình máy tính như AlphaGo đã giúp máy tính đánh bại con người trong các trò chơi như cờ vua, cờ vây được xây dựng dựa trên thuật toán này.

1.1.4. Các ứng dụng của Machine Learning

Học máy có ứng dụng rộng khắp trong các ngành khoa học/sản xuất, đặc biệt những ngành cần phân tích khối lượng dữ liệu khổng lồ. Một số ứng dụng thường thấy:

- Xử lý ngôn ngữ tự nhiên (Natural Language Processing): xử lý văn bản, giao tiếp người – máy,…
- Nhận dạng (Pattern Recognition): nhận dạng tiếng nói, chữ viết tay, vân tay, thị giác máy,…
- Tìm kiếm (Search Engine)
- Chẩn đoán trong y tế: phân tích ảnh X-quang, các hệ chuyên gia chẩn đoán tự động.
- Tin sinh học: phân loại chuỗi gene, quá trình hình thành gene/protein.
- Vật lý: phân tích ảnh thiên văn, tác động giữa các hạt,…
- Phát hiện gian lận tài chính (financial fraud): gian lận thẻ tín dụng.
- Phân tích thị trường chứng khoán (stock market analysis)
- Chơi trò chơi: tự động chơi cờ, hành động của các nhân vật ảo,…
- Robot: là tổng hợp của rất nhiều ngành khoa học, trong đó học máy tạo nên hệ thần kinh/bộ não của người máy.

1.2. Bài toán dự đoán loại lúa mì

1.2.1. Giới thiệu bài toán

Nhóm được kiểm tra bao gồm các hạt thuộc ba giống lúa mì khác nhau: Kama, Rosa và Canada, mỗi loại có 70 nguyên tố, được chọn ngẫu nhiên cho thí nghiệm. Hình dung chất lượng cao của cấu trúc hạt nhân bên trong được phát hiện bằng kỹ thuật X-quang mềm. Nó không phá hủy và chụp rẻ hơn đáng kể so với các kỹ thuật ảnh phức tạp khác như kính hiển vi quét hoặc công nghệ laser. Hình ảnh được ghi trên tấm

KODAK tia X 13x18 cm. Các nghiên cứu được thực hiện bằng cách sử dụng hạt lúa mì thu hoạch kết hợp có nguồn gốc từ các cánh đồng thí nghiệm, được khám phá tại Viện Vật lý nông nghiệp của Viện Hàn lâm Khoa học Ba Lan ở Lublin.

1.2.2. Dữ liệu

Dữ liệu cho bài toán này bao gồm bảy thông số hình học của hạt lúa mì đã được đo: diện tích hạt, chu vi hạt, độ nén hạt, chiều dài hạt, chiều rộng hạt, hệ số bất đối xứng và chiều dài rãnh hạt. Tất cả các tham số đều là giá trị thực.

1.2.3. Mục tiêu

Mục tiêu của bài toán là xây dựng một mô hình dự đoán loại lúa mì dựa trên các thông số thực tế. Mô hình này có thể giúp nhận biết và phân loại lúa mì tự động, từ đó hỗ trợ người nông dân trong việc quyết định về chăm sóc và quản lý loại lúa mì của họ.

1.2.4. Các bước thực hiện

a. Thu Thập Dữ Liệu

Thu thập bộ dữ liệu đa dạng về các loại lúa mì, bao gồm các thông số và nhãn chính xác cho từng loại.

b. Tiền Xử Lý Dữ Liệu

Chuẩn hóa các bộ dữ liệu, loại bỏ nhiễu, và thực hiện các bước tiền xử lý khác để làm cho dữ liệu phù hợp cho việc huấn luyện mô hình.

c. Xây Dựng Mô Hình

Sử dụng một mô hình học sâu, chẳng hạn như Decision Tree, để học từ dữ liệu và dự đoán loại lúa mì.

d. Huấn Luyện và Đánh Giá

Chia dữ liệu thành tập huấn luyện và tập kiểm thử. Huấn luyện mô hình trên tập huấn luyện và đánh giá hiệu suất trên tập kiểm thử để đảm bảo tính khả dụng và độ chính xác.

1.2.5. Ứng dụng thực tế

Khi mô hình đã được huấn luyện thành công, nó có thể tích hợp vào các hệ thống tự động trong nông nghiệp để tự động nhận diện và phân loại loại lúa mì. Điều này giúp nâng cao hiệu suất sản xuất và tiết kiệm thời gian cho người nông dân.

CHƯƠNG 2. MÔ HÌNH HỌC MÁY

2.1. Ý tưởng mô hình

Mô hình cây quyết định là một loại mô hình học máy được sử dụng để phân loại và dự đoán. Mô hình này hoạt động bằng cách xây dựng một cây có các nhánh đại diện cho các quyết định, và các lá cây đại diện cho các kết quả có thể xảy ra.

Ý tưởng của mô hình cây quyết định là phân tách dữ liệu thành các nhóm dựa trên các thuộc tính của dữ liệu đó. Các thuộc tính được chọn để phân tách dữ liệu là những thuộc tính có thể giúp phân biệt rõ ràng các nhóm dữ liệu.

Để xây dựng một cây quyết định, ta cần thực hiện các bước sau:

- Chọn một thuộc tính để phân tách dữ liệu.
- Chia dữ liệu thành các nhóm dựa trên thuộc tính đã chọn.
- Lặp lại bước 1 và 2 cho mỗi nhóm dữ liệu con.

Quá trình lặp lại này sẽ tiếp tục cho đến khi tất cả các dữ liệu được phân thành các nhóm riêng biệt.

Mỗi nút trong cây quyết định đại diện cho một quyết định. Các nhánh của nút đại diện cho các kết quả có thể xảy ra của quyết định đó. Các lá cây đại diện cho các kết quả có thể xảy ra cuối cùng của cây.

Để dự đoán kết quả cho một mẫu dữ liệu mới, ta cần bắt đầu từ gốc của cây và đi theo các nhánh cho đến khi đến một lá cây. Kết quả của lá cây đó sẽ là kết quả dự đoán cho mẫu dữ liệu mới.

Để dự đoán kết quả cho một mẫu dữ liệu mới ta cần bắt đầu từ gốc của cây và đi theo các nhánh cho đến khi đến một lá cây. Kết quả của lá cây đó sẽ là kết quả dự đoán cho mẫu dữ liệu mới

2.2. Phân tích mô hình

2.2.1. Biểu diễn cây quyết định

Cấu Trúc Cây: Mô hình cây quyết định có cấu trúc dạng cây với các nút và lá. Mỗi nút đại diện cho một quyết định dựa trên một đặc trưng cụ thể.

Chiều Sâu Cây: Chiều sâu của cây là số lượng tầng nút từ gốc đến lá. Chiều sâu này càng cao thì mô hình càng phức tạp.

2.2.2. Quyết định tại các nút

Đặc Trưng Quan Trọng: Mô hình cây quyết định sử dụng các đặc trưng quan trọng để chia nút. Các thuật toán thống kê, như Gini impurity, đo lường độ chia của dữ liệu dựa trên các đặc trưng.

Ngưỡng Chia Nút: Ngưỡng chia nút là giá trị đặc trưng mà mô hình sử dụng để phân tách dữ liệu tại mỗi nút.

2.2.3. Hiệu suất mô hình

Đánh Giá Trên Tập Kiểm Thử: Hiệu suất của mô hình được đánh giá bằng cách sử dụng tập kiểm thử để đảm bảo tính tổng quát và tránh overfitting hoặc underfitting.

Ma trận Confusion: Ma trận confusion giúp đo lường độ chính xác của mô hình và đánh giá khả năng phân loại cho từng lớp.

2.2.4. Diễn giải quyết định

Dễ Diễn Giải: Cây quyết định dễ diễn giải do cấu trúc tuyến tính của nó. Mỗi quyết định có thể được giải thích dựa trên luật đơn giản về giá trị đặc trưng.

2.2.5. Visualize cây

Biểu Đồ Cây: Sử dụng biểu đồ cây để hiển thị cấu trúc của cây quyết định. Các thư viện như Graphviz cung cấp công cụ để biểu diễn cây một cách trực quan.

2.2.6. Xử lý Overfitting và Underfitting

Kiểm Soát Overfitting: Giảm chiều sâu cây, tăng mức độ chia tối thiểu hoặc sử dụng pruning để kiểm soát overfitting.

Kiểm Soát Underfitting: Tăng chiều sâu cây hoặc điều chỉnh các tham số để cải thiện underfitting.

2.2.7. Ưu, nhược điểm của mô hình

Ưu điểm:

- Dễ hiểu và giải thích: Cấu trúc của cây quyết định rất dễ hiểu và giải thích. Điều này khiến cho mô hình này trở nên phổ biến trong các ứng dụng mà cần có sự giải thích rõ ràng về kết quả dự đoán, chẳng hạn như trong y khoa.
- Có thể áp dụng cho nhiều loại dữ liệu khác nhau: Mô hình cây quyết định có thể được áp dụng cho nhiều loại dữ liệu khác nhau, bao gồm dữ liệu số, dữ liệu phân loại và dữ liệu thời gian.
- Có thể xử lý dữ liệu có nhiễu: Mô hình cây quyết định có khả năng xử lý dữ liệu có nhiễu tốt hơn so với một số mô hình học máy khác.

Nhược điểm:

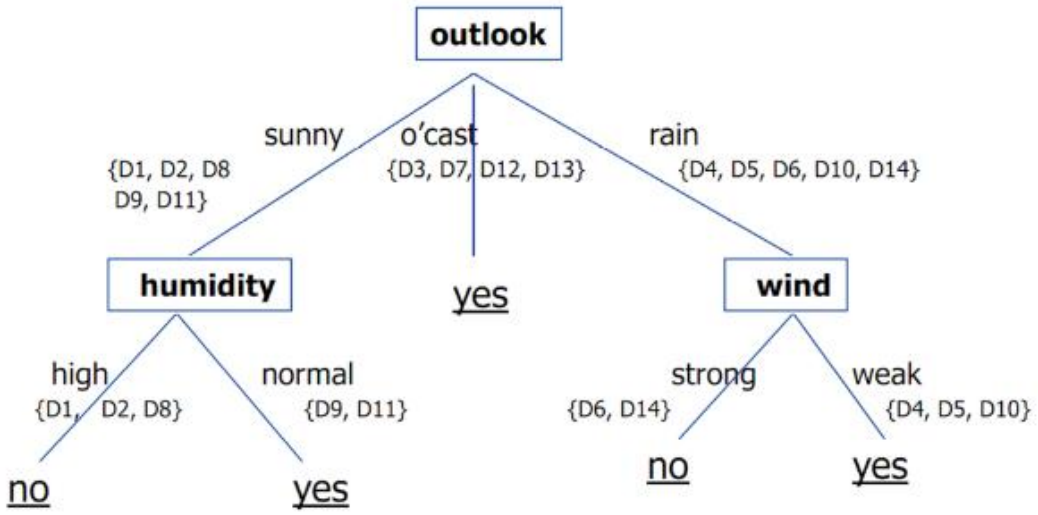
- Có thể quá phức tạp đối với dữ liệu có nhiều thuộc tính: Nếu dữ liệu có nhiều thuộc tính, cây quyết định có thể trở nên quá phức tạp và khó hiểu.
- Có thể bị quá khớp với dữ liệu huấn luyện: Mô hình cây quyết định có thể bị quá khớp với dữ liệu huấn luyện, dẫn đến việc dự đoán kém chính xác trên dữ liệu mới.

2.2.8. Ví dụ

Xem xét một ví dụ về một cây quyết định như sau:

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Từ bảng dữ liệu trên ta xây được cây quyết định như sau:



Cây quyết định của ví dụ trên có thể được giải thích như sau: các nút lá chứa các giá trị của thuộc tính phân lớp (thuộc tính "Play"). Các nút con tương ứng với các thuộc tính khác thuộc tính phân lớp. Nút gốc cũng được xem như một nút con đặc biệt, ở đây chính là thuộc tính "Outlook".

Các nhánh của cây từ một nút bất kỳ tương đương một phép so sánh có thể là so sánh bằng, so sánh khác, lớn hoặc nhỏ hơn,... nhưng kết quả các phép so sánh này bắt buộc phải thể hiện một giá trị logic (Đúng hoặc Sai) dựa trên một giá trị nào đó của thuộc tính của nút.

2.3. Cách thức đánh giá mô hình

2.3.1. Chia dữ liệu

Tập Huấn Luyện và Tập Kiểm Thử: Phân chia tập dữ liệu thành hai phần: một phần để huấn luyện mô hình (tập huấn luyện) và một phần để kiểm thử hiệu suất (tập kiểm thử).

2.3.2. Huấn luyện mô hình

Sử Dụng Tập Huấn Luyện: Đưa mô hình cây quyết định để huấn luyện trên tập dữ liệu huấn luyện.

Tinh chỉnh Tham Số: Điều chỉnh các tham số như chiều sâu cây, số lượng mẫu tối thiểu để chia nút để tối ưu hóa hiệu suất mô hình.

2.3.3. Kiểm thử trên tập kiểm thử

Dự Đoán Trên Tập Kiểm Thử: Sử dụng mô hình đã được huấn luyện để dự đoán trên tập kiểm thử, không được sử dụng trong quá trình huấn luyện.

Lấy Dự Đoán: Thu thập dự đoán từ mô hình cho mỗi mẫu trong tập kiểm thử.

2.3.4. Đánh giá hiệu suất

Độ Chính Xác (Accuracy): Tính tỷ lệ số lượng dự đoán đúng trên tổng số mẫu.

CHƯƠNG 3. CÀI ĐẶT MÔ HÌNH

3.1. Dữ liệu bài toán

Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1
14.38	14.21	0.8951	5.386	3.312	2.462	4.956	1
14.69	14.49	0.8799	5.563	3.259	3.586	5.219	1
14.11	14.1	0.8911	5.42	3.302	2.7	5	1
16.63	15.46	0.8747	6.053	3.465	2.04	5.877	1
16.44	15.25	0.888	5.884	3.505	1.969	5.533	1
15.26	14.85	0.8696	5.714	3.242	4.543	5.314	1
14.03	14.16	0.8796	5.438	3.201	1.717	5.001	1
13.89	14.02	0.888	5.439	3.199	3.986	4.738	1
13.78	14.06	0.8759	5.479	3.156	3.136	4.872	1
13.74	14.05	0.8744	5.482	3.114	2.932	4.825	1
14.59	14.28	0.8993	5.351	3.333	4.185	4.781	1
13.99	13.83	0.9183	5.119	3.383	5.234	4.781	1
15.69	14.75	0.9058	5.527	3.514	1.599	5.046	1
14.7	14.21	0.9153	5.205	3.466	1.767	4.649	1
12.72	13.57	0.8686	5.226	3.049	4.102	4.914	1
14.16	14.4	0.8584	5.658	3.129	3.072	5.176	1
14.11	14.26	0.8722	5.52	3.168	2.688	5.219	1
15.88	14.9	0.8988	5.618	3.507	0.7651	5.091	1
12.08	13.23	0.8664	5.099	2.936	1.415	4.961	1
15.01	14.76	0.8657	5.789	3.245	1.791	5.001	1
16.19	15.16	0.8849	5.833	3.421	0.903	5.307	1
13.02	13.76	0.8641	5.395	3.026	3.373	4.825	1
12.74	13.67	0.8564	5.395	2.956	2.504	4.869	1
14.11	14.18	0.882	5.541	3.221	2.754	5.038	1
13.45	14.02	0.8604	5.516	3.065	3.531	5.097	1
13.16	13.82	0.8662	5.454	2.975	0.8551	5.056	1
15.49	14.94	0.8724	5.757	3.371	3.412	5.228	1
14.09	14.41	0.8529	5.717	3.186	3.92	5.299	1
13.94	14.17	0.8728	5.585	3.15	2.124	5.012	1
15.05	14.68	0.8779	5.712	3.328	2.129	5.36	1
16.12	15	0.9	5.709	3.485	2.27	5.443	1
16.2	15.27	0.8734	5.826	3.464	2.823	5.527	1
17.08	15.38	0.9079	5.832	3.683	2.956	5.484	1
14.8	14.52	0.8823	5.656	3.288	3.112	5.309	1
14.28	14.17	0.8944	5.397	3.298	6.685	5.001	1
13.54	13.85	0.8871	5.348	3.156	2.587	5.178	1
13.5	13.85	0.8852	5.351	3.158	2.249	5.176	1

13.16	13.55	0.9009	5.138	3.201	2.461	4.783	1
15.5	14.86	0.882	5.877	3.396	4.711	5.528	1
15.11	14.54	0.8986	5.579	3.462	3.128	5.18	1
13.8	14.04	0.8794	5.376	3.155	1.56	4.961	1
15.36	14.76	0.8861	5.701	3.393	1.367	5.132	1
14.99	14.56	0.8883	5.57	3.377	2.958	5.175	1
14.79	14.52	0.8819	5.545	3.291	2.704	5.111	1
14.86	14.67	0.8676	5.678	3.258	2.129	5.351	1
14.43	14.4	0.8751	5.585	3.272	3.975	5.144	1
15.78	14.91	0.8923	5.674	3.434	5.593	5.136	1
14.49	14.61	0.8538	5.715	3.113	4.116	5.396	1
14.33	14.28	0.8831	5.504	3.199	3.328	5.224	1
14.52	14.6	0.8557	5.741	3.113	1.481	5.487	1
15.03	14.77	0.8658	5.702	3.212	1.933	5.439	1
14.46	14.35	0.8818	5.388	3.377	2.802	5.044	1
14.92	14.43	0.9006	5.384	3.412	1.142	5.088	1
15.38	14.77	0.8857	5.662	3.419	1.999	5.222	1
12.11	13.47	0.8392	5.159	3.032	1.502	4.519	1
11.42	12.86	0.8683	5.008	2.85	2.7	4.607	1
11.23	12.63	0.884	4.902	2.879	2.269	4.703	1
12.36	13.19	0.8923	5.076	3.042	3.22	4.605	1
13.22	13.84	0.868	5.395	3.07	4.157	5.088	1
12.78	13.57	0.8716	5.262	3.026	1.176	4.782	1
12.88	13.5	0.8879	5.139	3.119	2.352	4.607	1
14.34	14.37	0.8726	5.63	3.19	1.313	5.15	1
14.01	14.29	0.8625	5.609	3.158	2.217	5.132	1
14.37	14.39	0.8726	5.569	3.153	1.464	5.3	1
12.73	13.75	0.8458	5.412	2.882	3.533	5.067	1
17.63	15.98	0.8673	6.191	3.561	4.076	6.06	2
16.84	15.67	0.8623	5.998	3.484	4.675	5.877	2
17.26	15.73	0.8763	5.978	3.594	4.539	5.791	2
19.11	16.26	0.9081	6.154	3.93	2.936	6.079	2
16.82	15.51	0.8786	6.017	3.486	4.004	5.841	2
16.77	15.62	0.8638	5.927	3.438	4.92	5.795	2
17.32	15.91	0.8599	6.064	3.403	3.824	5.922	2
20.71	17.23	0.8763	6.579	3.814	4.451	6.451	2
18.94	16.49	0.875	6.445	3.639	5.064	6.362	2
17.12	15.55	0.8892	5.85	3.566	2.858	5.746	2
16.53	15.34	0.8823	5.875	3.467	5.532	5.88	2
18.72	16.19	0.8977	6.006	3.857	5.324	5.879	2
20.2	16.89	0.8894	6.285	3.864	5.173	6.187	2
19.57	16.74	0.8779	6.384	3.772	1.472	6.273	2
19.51	16.71	0.878	6.366	3.801	2.962	6.185	2
18.27	16.09	0.887	6.173	3.651	2.443	6.197	2
18.88	16.26	0.8969	6.084	3.764	1.649	6.109	2
18.98	16.66	0.859	6.549	3.67	3.691	6.498	2
21.18	17.21	0.8989	6.573	4.033	5.78	6.231	2

20.88	17.05	0.9031	6.45	4.032	5.016	6.321	2
20.1	16.99	0.8746	6.581	3.785	1.955	6.449	2
18.76	16.2	0.8984	6.172	3.796	3.12	6.053	2
18.81	16.29	0.8906	6.272	3.693	3.237	6.053	2
18.59	16.05	0.9066	6.037	3.86	6.001	5.877	2
18.36	16.52	0.8452	6.666	3.485	4.933	6.448	2
16.87	15.65	0.8648	6.139	3.463	3.696	5.967	2
19.31	16.59	0.8815	6.341	3.81	3.477	6.238	2
18.98	16.57	0.8687	6.449	3.552	2.144	6.453	2
18.17	16.26	0.8637	6.271	3.512	2.853	6.273	2
18.72	16.34	0.881	6.219	3.684	2.188	6.097	2
16.41	15.25	0.8866	5.718	3.525	4.217	5.618	2
17.99	15.86	0.8992	5.89	3.694	2.068	5.837	2
19.46	16.5	0.8985	6.113	3.892	4.308	6.009	2
19.18	16.63	0.8717	6.369	3.681	3.357	6.229	2
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	2
18.83	16.29	0.8917	6.037	3.786	2.553	5.879	2
18.85	16.17	0.9056	6.152	3.806	2.843	6.2	2
17.63	15.86	0.88	6.033	3.573	3.747	5.929	2
19.94	16.92	0.8752	6.675	3.763	3.252	6.55	2
18.55	16.22	0.8865	6.153	3.674	1.738	5.894	2
18.45	16.12	0.8921	6.107	3.769	2.235	5.794	2
19.38	16.72	0.8716	6.303	3.791	3.678	5.965	2
19.13	16.31	0.9035	6.183	3.902	2.109	5.924	2
19.14	16.61	0.8722	6.259	3.737	6.682	6.053	2
20.97	17.25	0.8859	6.563	3.991	4.677	6.316	2
19.06	16.45	0.8854	6.416	3.719	2.248	6.163	2
18.96	16.2	0.9077	6.051	3.897	4.334	5.75	2
19.15	16.45	0.889	6.245	3.815	3.084	6.185	2
18.89	16.23	0.9008	6.227	3.769	3.639	5.966	2
20.03	16.9	0.8811	6.493	3.857	3.063	6.32	2
20.24	16.91	0.8897	6.315	3.962	5.901	6.188	2
18.14	16.12	0.8772	6.059	3.563	3.619	6.011	2
16.17	15.38	0.8588	5.762	3.387	4.286	5.703	2
18.43	15.97	0.9077	5.98	3.771	2.984	5.905	2
15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
18.75	16.18	0.8999	6.111	3.869	4.188	5.992	2
18.65	16.41	0.8698	6.285	3.594	4.391	6.102	2
17.98	15.85	0.8993	5.979	3.687	2.257	5.919	2
20.16	17.03	0.8735	6.513	3.773	1.91	6.185	2
17.55	15.66	0.8991	5.791	3.69	5.366	5.661	2
18.3	15.89	0.9108	5.979	3.755	2.837	5.962	2
18.94	16.32	0.8942	6.144	3.825	2.908	5.949	2
15.38	14.9	0.8706	5.884	3.268	4.462	5.795	2
16.16	15.33	0.8644	5.845	3.395	4.266	5.795	2
15.56	14.89	0.8823	5.776	3.408	4.972	5.847	2
15.38	14.66	0.899	5.477	3.465	3.6	5.439	2

17.36	15.76	0.8785	6.145	3.574	3.526	5.971	2
15.57	15.15	0.8527	5.92	3.231	2.64	5.879	2
15.6	15.11	0.858	5.832	3.286	2.725	5.752	2
16.23	15.18	0.885	5.872	3.472	3.769	5.922	2
13.07	13.92	0.848	5.472	2.994	5.304	5.395	3
13.32	13.94	0.8613	5.541	3.073	7.035	5.44	3
13.34	13.95	0.862	5.389	3.074	5.995	5.307	3
12.22	13.32	0.8652	5.224	2.967	5.469	5.221	3
11.82	13.4	0.8274	5.314	2.777	4.471	5.178	3
11.21	13.13	0.8167	5.279	2.687	6.169	5.275	3
11.43	13.13	0.8335	5.176	2.719	2.221	5.132	3
12.49	13.46	0.8658	5.267	2.967	4.421	5.002	3
12.7	13.71	0.8491	5.386	2.911	3.26	5.316	3
10.79	12.93	0.8107	5.317	2.648	5.462	5.194	3
11.83	13.23	0.8496	5.263	2.84	5.195	5.307	3
12.01	13.52	0.8249	5.405	2.776	6.992	5.27	3
12.26	13.6	0.8333	5.408	2.833	4.756	5.36	3
11.18	13.04	0.8266	5.22	2.693	3.332	5.001	3
11.36	13.05	0.8382	5.175	2.755	4.048	5.263	3
11.19	13.05	0.8253	5.25	2.675	5.813	5.219	3
11.34	12.87	0.8596	5.053	2.849	3.347	5.003	3
12.13	13.73	0.8081	5.394	2.745	4.825	5.22	3
11.75	13.52	0.8082	5.444	2.678	4.378	5.31	3
11.49	13.22	0.8263	5.304	2.695	5.388	5.31	3
12.54	13.67	0.8425	5.451	2.879	3.082	5.491	3
12.02	13.33	0.8503	5.35	2.81	4.271	5.308	3
12.05	13.41	0.8416	5.267	2.847	4.988	5.046	3
12.55	13.57	0.8558	5.333	2.968	4.419	5.176	3
11.14	12.79	0.8558	5.011	2.794	6.388	5.049	3
12.1	13.15	0.8793	5.105	2.941	2.201	5.056	3
12.44	13.59	0.8462	5.319	2.897	4.924	5.27	3
12.15	13.45	0.8443	5.417	2.837	3.638	5.338	3
11.35	13.12	0.8291	5.176	2.668	4.337	5.132	3
11.24	13	0.8359	5.09	2.715	3.521	5.088	3
11.02	13	0.8189	5.325	2.701	6.735	5.163	3
11.55	13.1	0.8455	5.167	2.845	6.715	4.956	3
11.27	12.97	0.8419	5.088	2.763	4.309	5	3
11.4	13.08	0.8375	5.136	2.763	5.588	5.089	3
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	3
10.8	12.57	0.859	4.981	2.821	4.773	5.063	3
11.26	13.01	0.8355	5.186	2.71	5.335	5.092	3
10.74	12.73	0.8329	5.145	2.642	4.702	4.963	3
11.48	13.05	0.8473	5.18	2.758	5.876	5.002	3
12.21	13.47	0.8453	5.357	2.893	1.661	5.178	3
11.41	12.95	0.856	5.09	2.775	4.957	4.825	3
12.46	13.41	0.8706	5.236	3.017	4.987	5.147	3
12.19	13.36	0.8579	5.24	2.909	4.857	5.158	3

11.65	13.07	0.8575	5.108	2.85	5.209	5.135	3
12.89	13.77	0.8541	5.495	3.026	6.185	5.316	3
11.56	13.31	0.8198	5.363	2.683	4.062	5.182	3
11.81	13.45	0.8198	5.413	2.716	4.898	5.352	3
10.91	12.8	0.8372	5.088	2.675	4.179	4.956	3
11.23	12.82	0.8594	5.089	2.821	7.524	4.957	3
10.59	12.41	0.8648	4.899	2.787	4.975	4.794	3
10.93	12.8	0.839	5.046	2.717	5.398	5.045	3
11.27	12.86	0.8563	5.091	2.804	3.985	5.001	3
11.87	13.02	0.8795	5.132	2.953	3.597	5.132	3
10.82	12.83	0.8256	5.18	2.63	4.853	5.089	3
12.11	13.27	0.8639	5.236	2.975	4.132	5.012	3
12.8	13.47	0.886	5.16	3.126	4.873	4.914	3
12.79	13.53	0.8786	5.224	3.054	5.483	4.958	3
13.37	13.78	0.8849	5.32	3.128	4.67	5.091	3
12.62	13.67	0.8481	5.41	2.911	3.306	5.231	3
12.76	13.38	0.8964	5.073	3.155	2.828	4.83	3
12.38	13.44	0.8609	5.219	2.989	5.472	5.045	3
12.67	13.32	0.8977	4.984	3.135	2.3	4.745	3
11.18	12.72	0.868	5.009	2.81	4.051	4.828	3
12.7	13.41	0.8874	5.183	3.091	8.456	5	3
12.37	13.47	0.8567	5.204	2.96	3.919	5.001	3
12.19	13.2	0.8783	5.137	2.981	3.631	4.87	3
11.23	12.88	0.8511	5.14	2.795	4.325	5.003	3
13.2	13.66	0.8883	5.236	3.232	8.315	5.056	3
11.84	13.21	0.8521	5.175	2.836	3.598	5.044	3
12.3	13.34	0.8684	5.243	2.974	5.637	5.063	3

3.2. Hiển thị dữ liệu

3.2.1. Mã chương trình

```
import pandas as pd
data = pd.read_csv("seeds.csv")
columns = ['Dien Tich', 'Chu Vi', 'Do Nen', 'Chieu Dai', 'Chieu Rong','Hs BDX','Dai Ranh Hat','Loai Hat']
print(data)
```

```
     Dien Tich  Chu Vi  Do Nen  Chieu Dai  Chieu Rong  Hs BDX  Dai Ranh Hat  Loai Hat
0        15.26   14.84  0.8710      5.763       3.312   2.221         5.220         1
1        14.88   14.57  0.8811      5.554       3.333   1.018         4.956         1
2        14.29   14.09  0.9050      5.291       3.337   2.699         4.825         1
3        13.84   13.94  0.8955      5.324       3.379   2.259         4.805         1
4        16.14   14.99  0.9034      5.658       3.562   1.355         5.175         1
..         ...     ...     ...        ...         ...     ...           ...       ...
205      12.19   13.20  0.8783      5.137       2.981   3.631         4.870         3
206      11.23   12.88  0.8511      5.140       2.795   4.325         5.003         3
207      13.20   13.66  0.8883      5.236       3.232   8.315         5.056         3
208      11.84   13.21  0.8521      5.175       2.836   3.598         5.044         3
209      12.30   13.34  0.8684      5.243       2.974   5.637         5.063         3

[210 rows x 8 columns]
```

3.2.2. Kết quả

3.3. Cài đặt mô hình cây quyết định bằng Python (sklearn)

3.3.1 Mã chương trình

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
import seaborn as sns
from scipy import stats
import warnings

data = pd.read_csv("seeds.csv")
columns = ['Dien Tich', 'Chu Vi', 'Do Nen', 'Chieu Dai', 'Chieu
Rong','Hs BDX','Dai Ranh Hat','Loai Hat']
print(data)

from pandas.api.types import is_numeric_dtype
for col in data.columns:
    if is_numeric_dtype(data[col]):
        print('%s:' % (col))
        print('\t Mean = %.2f' % data[col].mean())
        print('\t Median = %.2f' % data[col].median())
        print('\t Mode = %.2f' % data[col].mode().values[0])
        print('\t Variance = %.2f' % data[col].var())
        print('\t Standard deviation = %.2f' % data[col].std())

# Tắt cảnh báo về feature names
warnings.filterwarnings("ignore", category=UserWarning)

sns.distplot(data['Dien Tich'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Dien Tich')
plt.xlabel('Dien Tich')
plt.ylabel('Tần suất')
plt.show()

sns.distplot(data['Chu Vi'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Chu Vi')
plt.xlabel('Chu Vi')
plt.ylabel('Tần suất')
plt.show()

sns.distplot(data['Do Nen'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Do Nen')
plt.xlabel('Do Nen')
plt.ylabel('Tần suất')
plt.show()

sns.distplot(data['Chieu Dai'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Chieu Dai')
plt.xlabel('Chieu Dai')
plt.ylabel('Tần suất')
plt.show()
```

```
sns.distplot(data['Chieu Rong'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Chieu Rong')
plt.xlabel('Chieu Rong')
plt.ylabel('Tần suất')
plt.show()

sns.distplot(data['Hs BDX'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Hs BDX')
plt.xlabel('Hs BDX')
plt.ylabel('Tần suất')
plt.show()

sns.distplot(data['Dai Ranh Hat'], kde=False, color='#1192e8')
plt.title('Biểu đồ histogram cho cột Dai Ranh Hat')
plt.xlabel('Dai Ranh Hat')
plt.ylabel('Tần suất')
plt.show()

sns.set_theme(style="darkgrid")
sns.countplot(x="Loai Hat", data=data)
plt.title('Biểu đồ histogram cho cột Loai Hat')
plt.xlabel('Loai Hat')
plt.ylabel('Tần suất')
plt.show()


# Kiểm tra giá trị ngoại lai
z_scores = stats.zscore(data['Dien Tich'])
threshold = 2
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Dien Tich không có giá trị ngoại lai")
else:
    print("Cột Dien Tich có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Chu Vi'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Chu Vi không có giá trị ngoại lai")
else:
    print("Cột Chu Vi có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Do Nen'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Do Nen không có giá trị ngoại lai")
else:
```

```
    print("Cột Do Nen có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Chieu Dai'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Chieu Dai không có giá trị ngoại lai")
else:
    print("Cột Chieu Dai có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Chieu Rong'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Chieu Rong không có giá trị ngoại lai")
else:
    print("Cột Chieu Rong có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Hs BDX'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Hs BDX không có giá trị ngoại lai")
else:
    print("Cột Hs BDX có giá trị ngoại lai")
    print(outliers)

z_scores = stats.zscore(data['Dai Ranh Hat'])
outliers = data[abs(z_scores) > threshold]
if outliers.empty:
    print("Cột Dai Ranh Hat không có giá trị ngoại lai")
else:
    print("Cột Dai Ranh Hat có giá trị ngoại lai")
    print(outliers)


#------------
#Đánh giá tỷ lệ và tính chất của dữ liệu bị thiếu
data.columns[data.isnull().any()]
data.isnull().sum()
print("Số lượng dữ liệu bị thiếu của từng cột là:\n",data.isnull().sum())

x = data[['Dien Tich', 'Chu Vi', 'Do Nen', 'Chieu Dai', 'Chieu Rong','Hs
BDX','Dai Ranh Hat']].values
y = data["Loai Hat"]


#Biểu đồ
sns.boxplot(data=data[columns])
plt.xlabel('Tên cột')
```

```
plt.ylabel('Giá trị')
plt.title('Biểu đồ Boxplot')
plt.show()

# Vẽ biểu đồ heatmap của ma trận tương quan
correlation_matrix = data.corr()
plt.figure(figsize=(10,10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Biểu đồ thể hiện ma trận tương quan')
plt.show()

#Xây dựng cây quyết định
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,
random_state=3)

#Mô hình hóa
dt = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
dt.fit(x_train,y_train)
#Dự đoán
pt = dt.predict(x_test)
```

3.3.2 Kết quả

```
Dien Tich:
        Mean = 14.85
        Median = 14.36
        Mode = 11.23
        Variance = 8.47
        Standard deviation = 2.91
Chu Vi:
        Mean = 14.56
        Median = 14.32
        Mode = 13.47
        Variance = 1.71
        Standard deviation = 1.31
Do Nen:
        Mean = 0.87
        Median = 0.87
        Mode = 0.88
        Variance = 0.00
        Standard deviation = 0.02
Chieu Dai:
        Mean = 5.63
        Median = 5.52
        Mode = 5.24
        Variance = 0.20
        Standard deviation = 0.44
Hs BDX:
        Mean = 3.70
        Median = 3.60
        Mode = 2.13
        Variance = 2.26
        Standard deviation = 1.50
Dai Ranh Hat:
        Mean = 5.41
        Median = 5.22
        Mode = 5.00
        Variance = 0.24
        Standard deviation = 0.49
Loai Hat:
        Mean = 2.00
        Median = 2.00
        Mode = 1.00
        Variance = 0.67
        Standard deviation = 0.82
```

Cột Dien Tich có giá trị ngoại lai

	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
77	20.71	17.23	0.8763	6.579	3.814	4.451	6.451	2
88	21.18	17.21	0.8989	6.573	4.033	5.780	6.231	2
89	20.88	17.05	0.9031	6.450	4.032	5.016	6.321	2
114	20.97	17.25	0.8859	6.563	3.991	4.677	6.316	2

Cột Chu Vi có giá trị ngoại lai

	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
77	20.71	17.23	0.8763	6.579	3.814	4.451	6.451	2
88	21.18	17.21	0.8989	6.573	4.033	5.780	6.231	2
114	20.97	17.25	0.8859	6.563	3.991	4.677	6.316	2

Cột Do Nen có giá trị ngoại lai

	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
16	13.99	13.83	0.9183	5.119	3.383	5.234	4.781	1
145	11.21	13.13	0.8167	5.279	2.687	6.169	5.275	3
149	10.79	12.93	0.8107	5.317	2.648	5.462	5.194	3
157	12.13	13.73	0.8081	5.394	2.745	4.825	5.220	3
158	11.75	13.52	0.8082	5.444	2.678	4.378	5.310	3
170	11.02	13.00	0.8189	5.325	2.701	6.735	5.163	3
174	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	3
185	11.56	13.31	0.8198	5.363	2.683	4.062	5.182	3
186	11.81	13.45	0.8198	5.413	2.716	4.898	5.352	3

Cột Chieu Dai có giá trị ngoại lai

	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
77	20.71	17.23	0.8763	6.579	3.814	4.451	6.451	2
87	18.98	16.66	0.8590	6.549	3.670	3.691	6.498	2
88	21.18	17.21	0.8989	6.573	4.033	5.780	6.231	2
90	20.10	16.99	0.8746	6.581	3.785	1.955	6.449	2
94	18.36	16.52	0.8452	6.666	3.485	4.933	6.448	2
108	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	2
114	20.97	17.25	0.8859	6.563	3.991	4.677	6.316	2
128	20.16	17.03	0.8735	6.513	3.773	1.910	6.185	2

Cột Chieu Rong có giá trị ngoại lai

	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
88	21.18	17.21	0.8989	6.573	4.033	5.780	6.231	2
89	20.88	17.05	0.9031	6.450	4.032	5.016	6.321	2

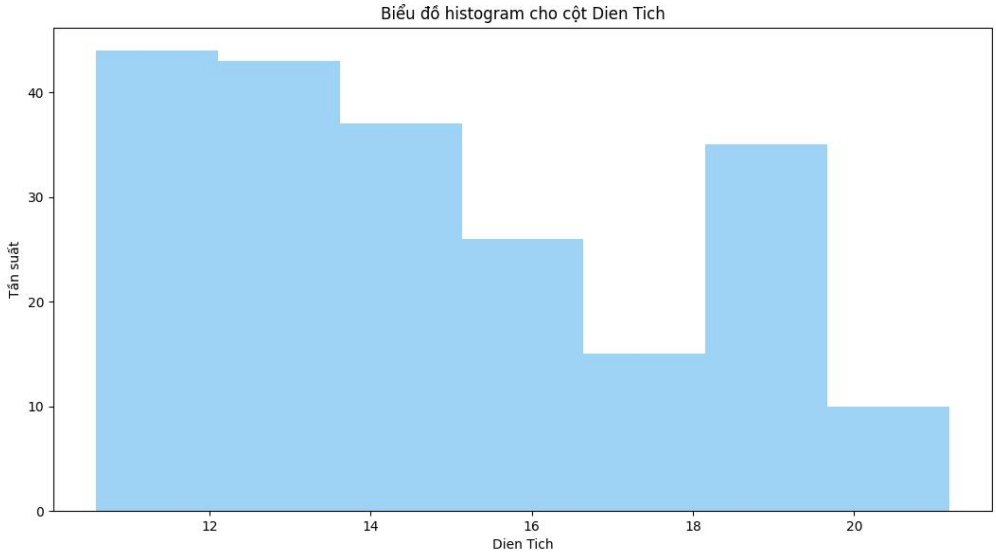
Cột Hs BDX có giá trị ngoại lai

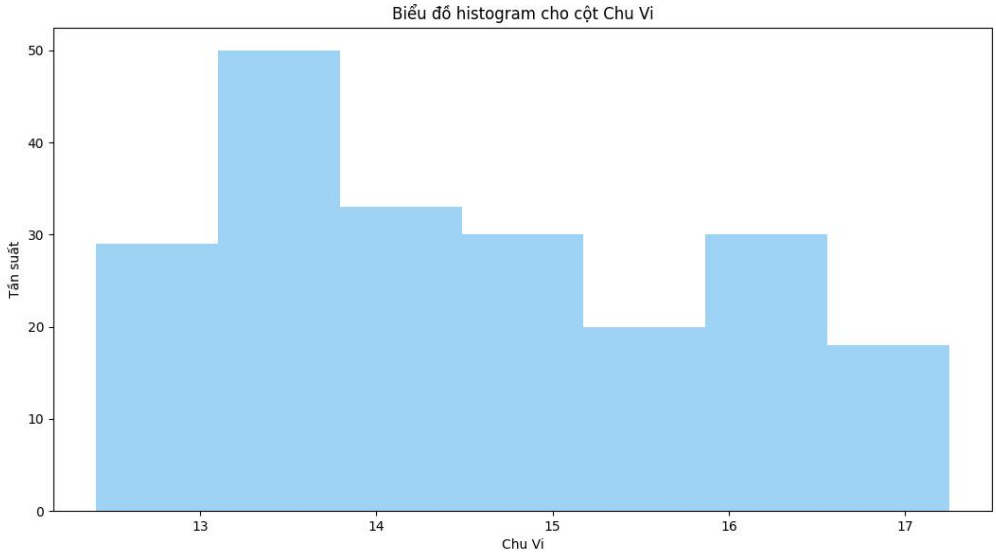
	Dien Tich	Chu Vi	Do Nen	Chieu Dai	Chieu Rong	Hs BDX	Dai Ranh Hat	Loai Hat
141	13.32	13.94	0.8613	5.541	3.073	7.035	5.440	3
151	12.01	13.52	0.8249	5.405	2.776	6.992	5.270	3
170	11.02	13.00	0.8189	5.325	2.701	6.735	5.163	3
171	11.55	13.10	0.8455	5.167	2.845	6.715	4.956	3
188	11.23	12.82	0.8594	5.089	2.821	7.524	4.957	3
90	20.10	16.99	0.8746	6.581	3.785	1.955	6.449	2
94	18.36	16.52	0.8452	6.666	3.485	4.933	6.448	2
97	18.98	16.57	0.8687	6.449	3.552	2.144	6.453	2
108	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	2

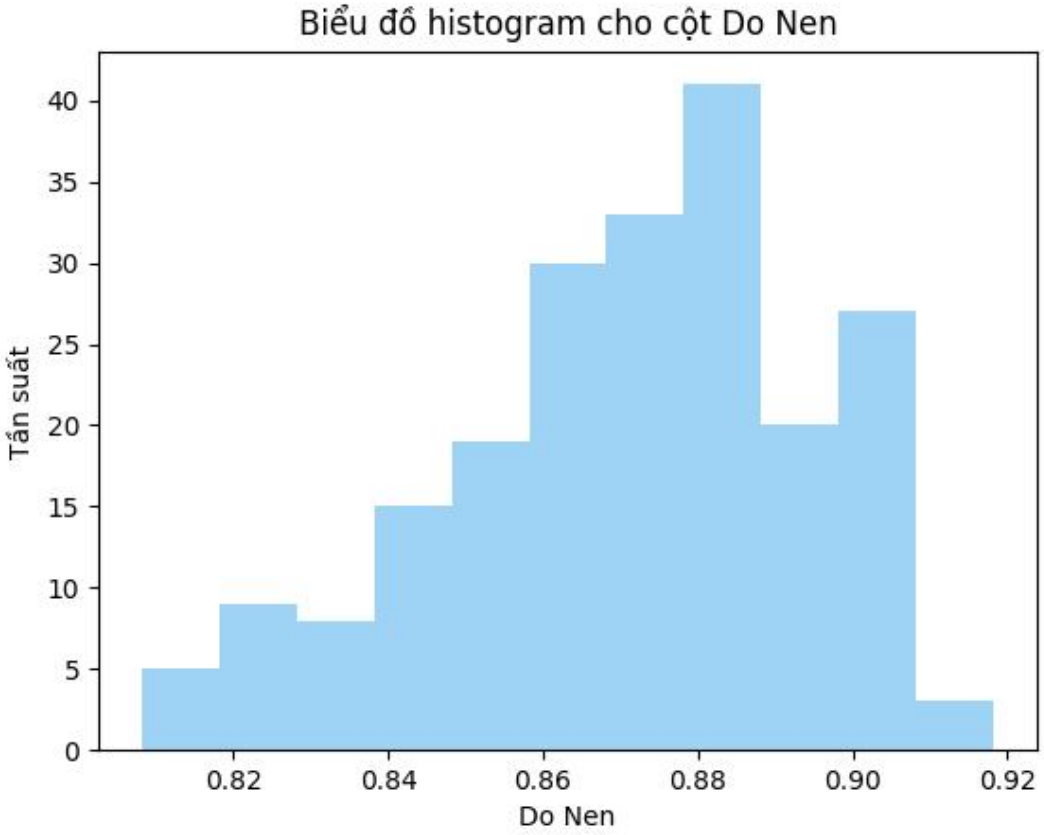
Số lượng dữ liệu bị thiếu của từng cột là:

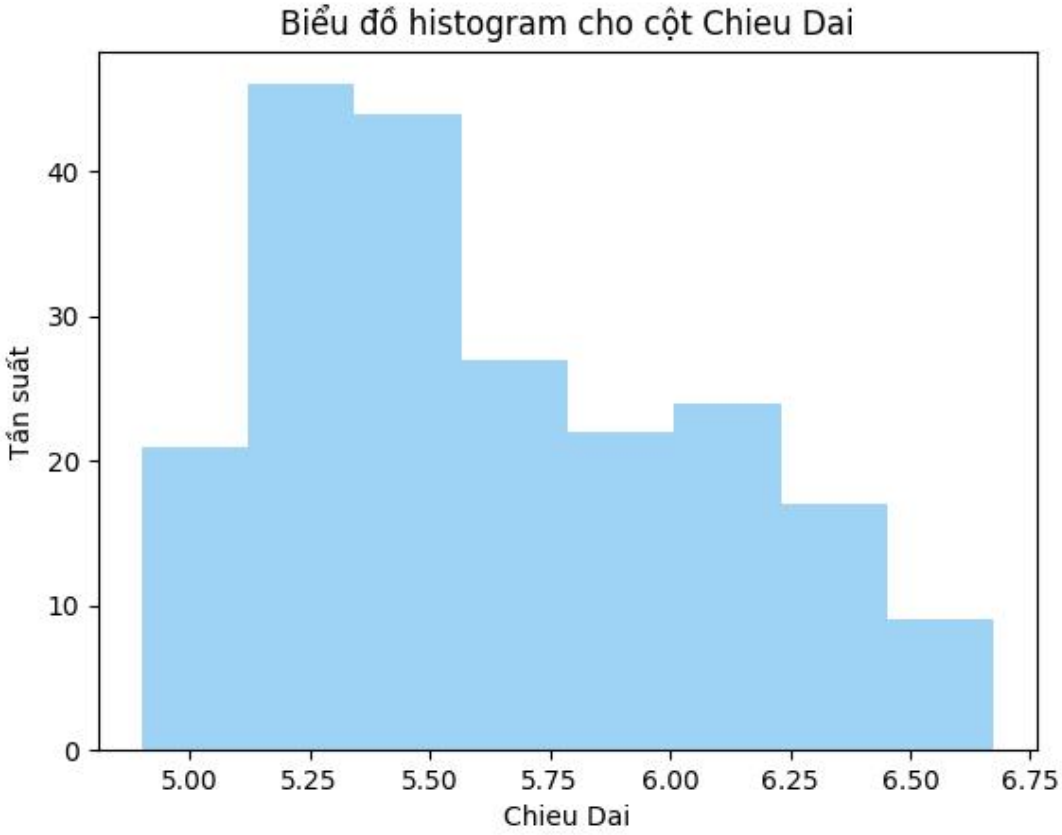
```
 Dien Tich       0
Chu Vi           0
Do Nen           0
Chieu Dai        0
Chieu Rong       0
Hs BDX           0
Dai Ranh Hat     0
Loai Hat         0
dtype: int64
```

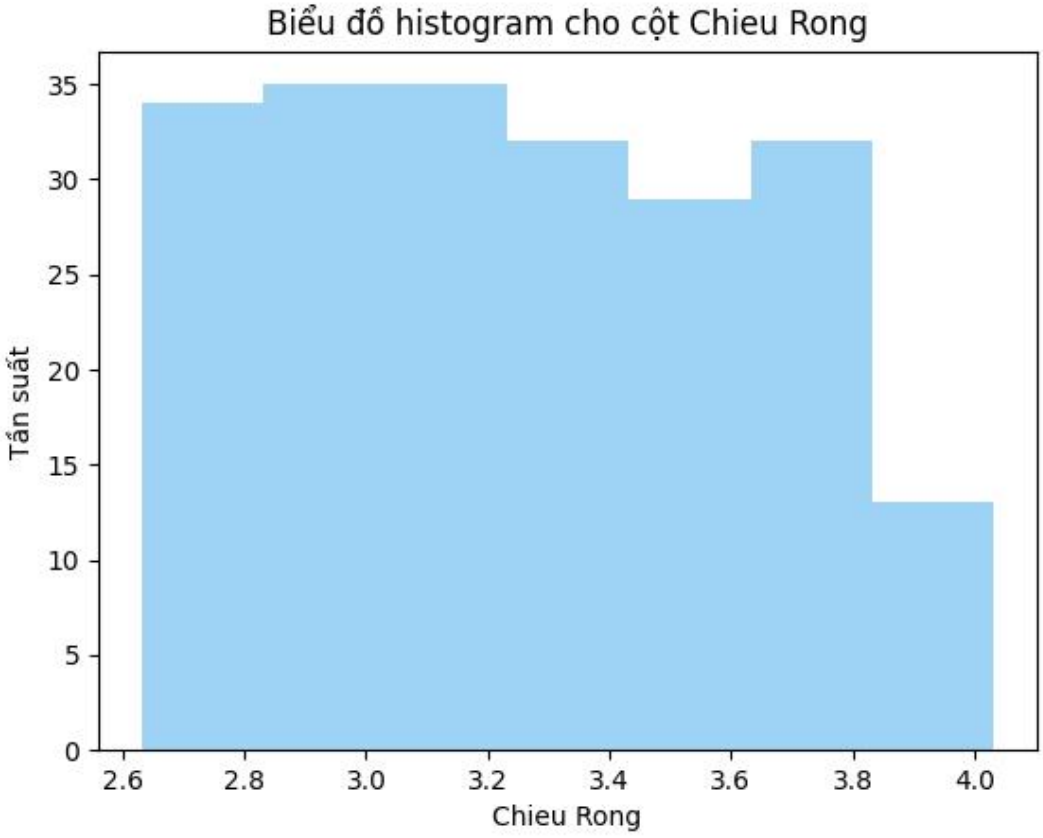
3.3.3 Các biểu đồ

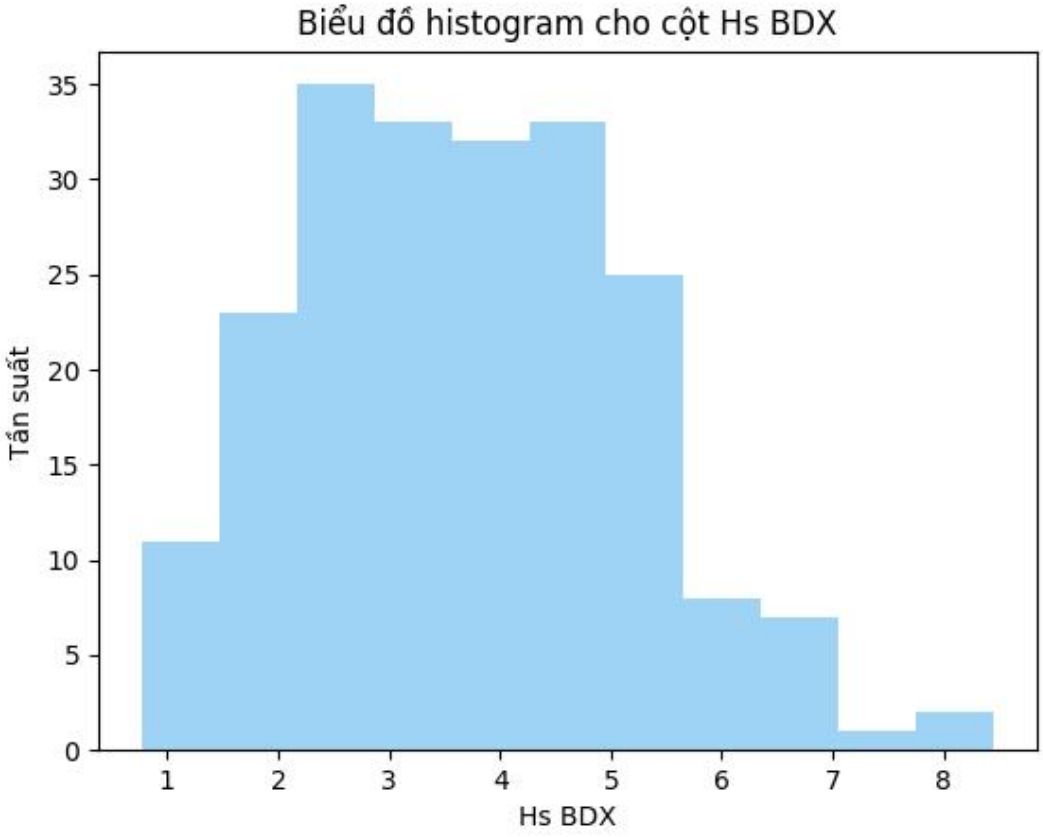

Biểu đồ histogram cho cột Dien Tich
Tần suất
Dien Tich
0
10
20
30
40
12
14
16
18
20

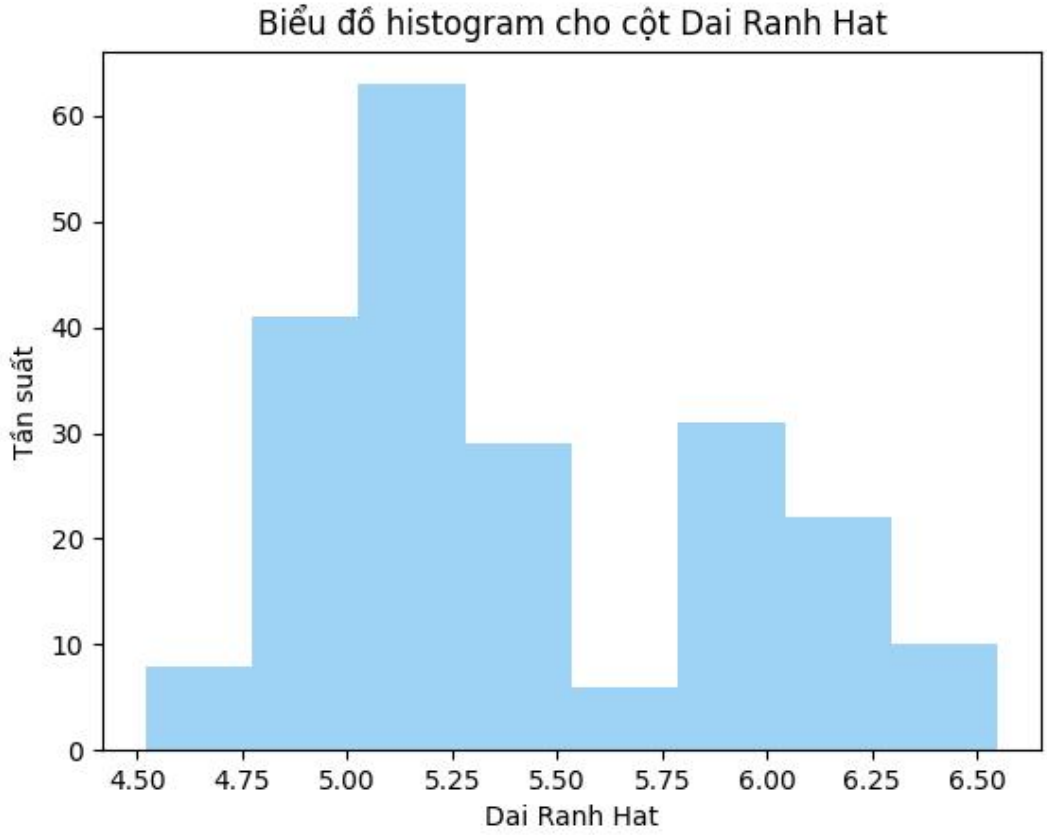

Biểu đồ histogram cho cột Chu Vi
Tần suất
Chu Vi
0
10
20
30
40
50
13
14
15
16
17

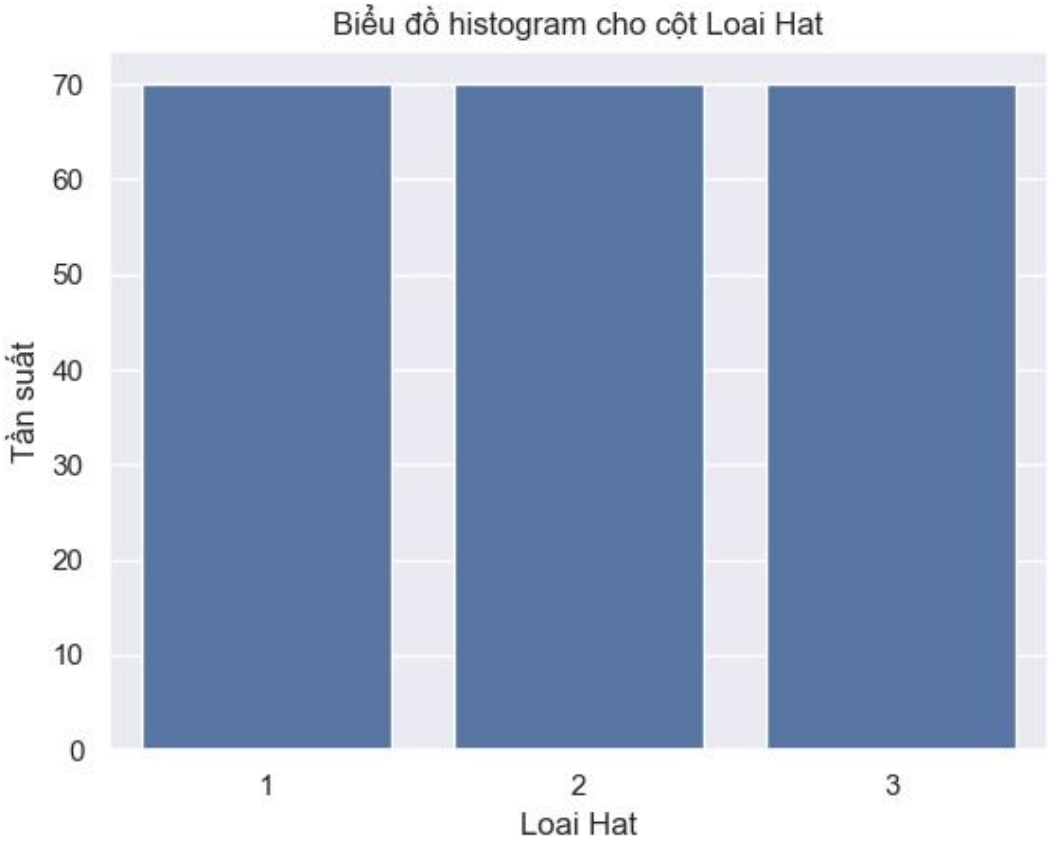

Biểu đồ histogram cho cột Do Nen
Tần suất
0
5
10
15
20
25
30
35
40
0.82
0.84
0.86
0.88
0.90
0.92
Do Nen

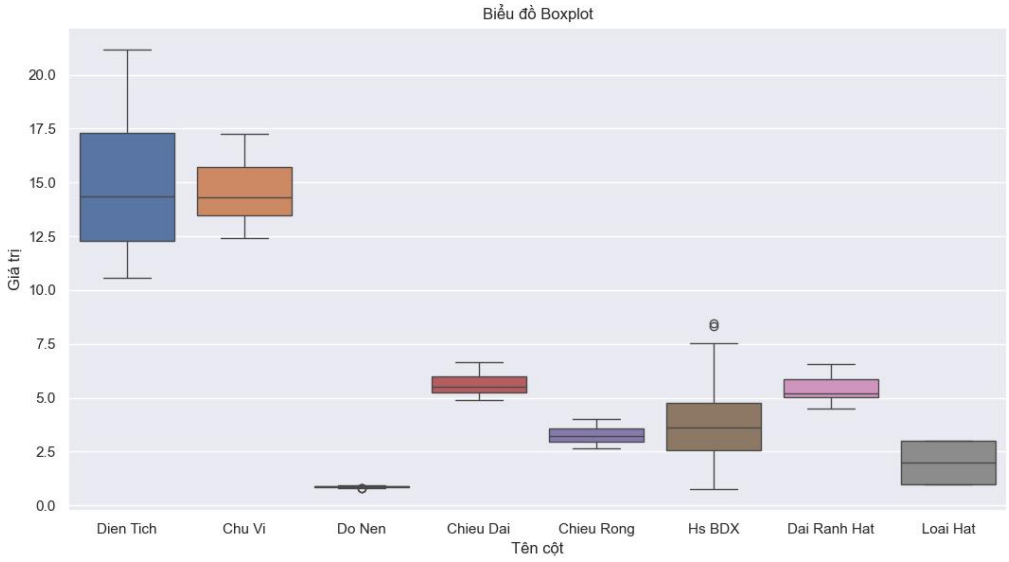

Biểu đồ histogram cho cột Chieu Dai
Tần suất
0
10
20
30
40
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
Chieu Dai

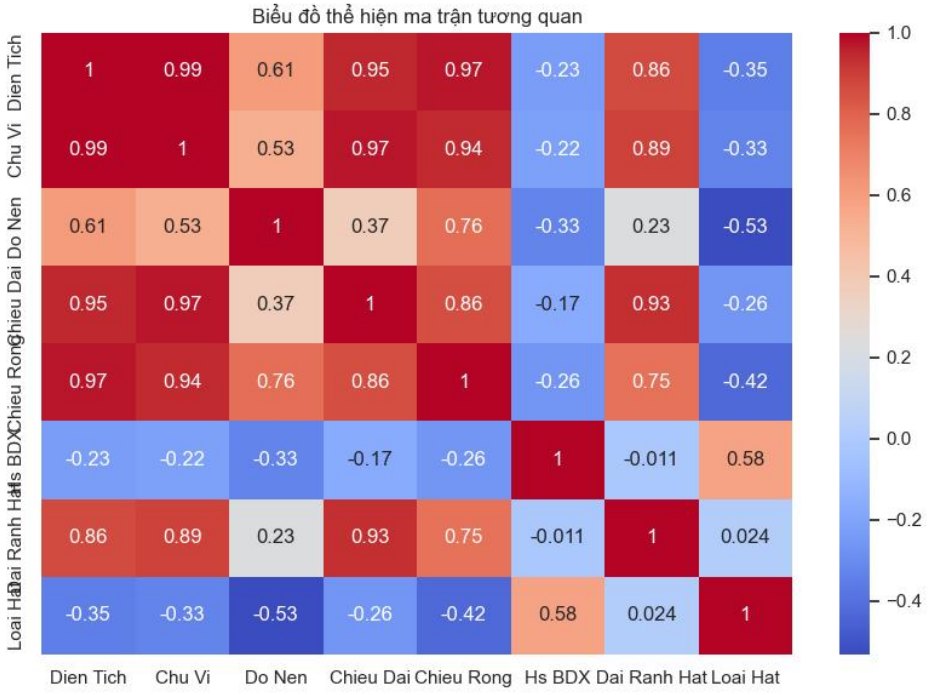

Biểu đồ histogram cho cột Chieu Rong
Tần suất
Chieu Rong
0
5
10
15
20
25
30
35
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0


Biểu đồ histogram cho cột Hs BDX
Tần suất
Hs BDX
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8


Biểu đồ histogram cho cột Dai Ranh Hat
Tần suất
0
10
20
30
40
50
60
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
Dai Ranh Hat


Biểu đồ histogram cho cột Loai Hat
Tần suất
0
10
20
30
40
50
60
70
1
2
3
Loai Hat





3.4. Đánh giá độ chính xác của mô hình

3.4.1 Mã chương trình

```
#Đánh giá
from sklearn import metrics
print("Độ chính xác của cây quyết định : ", metrics.accuracy_score(y_test,pt))
```

3.4.2 Kết quả

```
Độ chính xác của cây quyết định :  0.9365079365079365
```

3.4.3 Kết luận

Mô hình đã đạt được một độ chính xác cao là xấp xỉ 0.94, thể hiện khả năng dự đoán chính xác đáng kể trên tập dữ liệu kiểm thử. Kết quả này là một dấu hiệu tích cực về hiệu suất của mô hình trong việc phân loại dữ liệu. Độ chính xác 0.94 có nghĩa là mô hình dự đoán đúng khoảng 94% trên tổng số lượng dự đoán.

TỔNG KẾT

1. Các nhiệm vụ đã thực hiện

- **Bài Toán Dự Đoán Loại Lúa Mì**

  Mô tả bài toán dự đoán loại lúa mì dựa trên nhiều đặc trưng.

- **Mô Hình Cây Quyết Định**

  Giới thiệu mô hình cây quyết định với ý tưởng cơ bản về quyết định dựa trên đặc trưng để phân loại dữ liệu.

- **Triển Khai Mô Hình**

  Lựa chọn và cài đặt mô hình sử dụng thư viện học máy scikit-learn trong ngôn ngữ lập trình Python.

- **Huấn Luyện và Đánh Giá**

  Mô tả quy trình chia dữ liệu, huấn luyện mô hình trên tập dữ liệu huấn luyện, và đánh giá hiệu suất trên tập kiểm thử.
  Sử dụng các phương pháp đánh giá như độ chính xác để đo lường khả năng dự đoán của mô hình.

- **Kết Quả và Hiệu Suất Mô Hình**

  Mô hình đã đạt được một độ chính xác cao là 0.94, chiếm 94% tỷ lệ dự đoán đúng trên tổng số lượng dự đoán.
  Kết quả này là một bước tiến tích cực và cho thấy khả năng xuất sắc của mô hình trong việc phân loại loại lúa mì.

- **Báo cáo**

  Viết báo cáo lại quá trình làm việc cũng như giới thiệu về chương trình.

2. Ưu điểm và nhược điểm của chương trình

- **Ưu điểm**

  Quan Trọng Cho Nông Nghiệp: Bài toán dự đoán loại lúa mì có ý nghĩa quan trọng trong lĩnh vực nông nghiệp. Việc chính xác xác định loại lúa mì có thể hỗ trợ những quyết định liên quan đến chăm sóc cây trồng, sử dụng nguồn tài nguyên và tối ưu hóa sản xuất.
  Ứng Dụng Học Máy Trong Nông Nghiệp: Bài toán là một ví dụ tốt về ứng dụng của học máy trong lĩnh vực nông nghiệp.

Việc sử dụng mô hình cây quyết định giúp tự động hóa quá trình phân loại và dự đoán loại lúa mì.

Tính Ứng Dụng Rộng Rãi: Kỹ thuật và mô hình phát triển trong bài toán có thể mở rộng để áp dụng cho nhiều loại cây trồng khác nhau, mở cánh cửa cho ứng dụng trong các lĩnh vực khác của nông nghiệp.

- **Nhược điểm**

Phụ Thuộc vào Dữ Liệu: Hiệu suất của mô hình đặc biệt phụ thuộc vào chất lượng và đại diện của dữ liệu huấn luyện. Nếu dữ liệu không đủ đại diện hoặc có nhiễu, mô hình có thể không dự đoán chính xác.

Khả Năng Mất Cân Bằng Lớp: Nếu sự phân bố giữa các loại lúa mì trong tập dữ liệu là không đồng đều, mô hình có thể chịu ảnh hưởng của mất cân bằng lớp và dự đoán không chính xác cho các lớp thiểu số.

Khả Năng Overfitting hoặc Underfitting: Nếu không tinh chỉnh thích hợp, mô hình cây quyết định có thể dễ bị overfitting (quá khớp) hoặc underfitting (không khớp đủ) trên dữ liệu, ảnh hưởng đến khả năng tổng quát hóa của nó.

Khả Năng Chưa Đối Phó với Biến Đổi Môi Trường: Mô hình có thể không đối phó tốt với biến đổi môi trường như điều kiện thời tiết hoặc biến đổi khác trong quá trình sản xuất lúa mì.

TÀI LIỆU THAM KHẢO

[1] Bùi Việt Hà, Python cơ bản, NXB Đại học Quốc Gia Hà Nội, 2023.

[2] TS. Nguyễn Văn Hậu , TS. Phạm Minh Chuẩn, TS. Nguyễn Văn Quyết, Giáo trình học máy cơ bản, NXB Khoa học và Kỹ Thuật, 2022.