**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC XÃ HỘI VÀ NHÂN VĂN TPHCM**

**KHOA THƯ VIỆN - THÔNG TIN HỌC**

------------------------------<<>>------------------------------



**ĐỒ ÁN CUỐI KỲ**

**PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ CỔ PHIẾU BẰNG MÔ HÌNH DỰ BÁO**

Môn: Phân tích dữ liệu cho quản lý

Giảng viên : Trần Đình Anh Huy

Nhóm 1 – Lớp Quản lý thông tin B

|  |  |  |
| --- | --- | --- |
| STT | Họ và tên | Mã số sinh viên |
| 1 | Hoàng Xuân Quốc | 2156210125 |
| 2 | Đặng Hoàng Chiến | 2156210095 |
| 3 | Nguyễn Viết Đức | 2156210100 |

**MỤC LỤC**

1. TỔNG QUAN
2. Giới thiệu nhóm /lý do chọn dự án

Học máy (ML - Machine Learning) đã trở thành thành phần công nghệ cốt lõi trong nhiều trường hợp sử dụng từ xử lý ngôn ngữ tự nhiên và thị giác máy tính đến phát hiện gian lận, dự báo nhu cầu, đề xuất sản phẩm, bảo trì phòng ngừa và xử lý tài liệu. Với những kiến thức đã học được trong môn “ Lưu trữ và khai thác hệ thống” của thầy “ NT Công” chúng em quyết định chọn đề tài là một trong những thuật toán phổ biến trong lĩnh vực học máy là K-Nearest Neighbors (KNN), một thuật toán phân loại dựa trên việc đo khoảng cách giữa các điểm dữ liệu. KNN là thuật toán học máy đóng vai trò quan trọng trong việc phân loại văn bản, nhận dạng hình ảnh và nhiều tác vụ khác trong lĩnh vực xử lý ngôn ngữ tự nhiên và thị giác máy tính. Nó là ngôn ngữ học máy trong quy mô lớn. Để tận dụng cần phải tiêu chuẩn hóa quy trình phát triển ML hiện đại trên toàn doanh nghiệp.

Điều quan trọng hơn hết đến việc lựa chọn đề tài KNN cho phép các nhà phát triển và nhà khoa học dữ liệu ở mọi cấp độ kỹ năng ML truy cập và quản lý tài nguyên một cách hiệu quả, đồng thời duy trì chi phí thấp. Nhờ vào khả năng phân loại chính xác và tính linh hoạt của nó, KNN là một công cụ mạnh mẽ trong việc xử lý ngôn ngữ tự nhiên và thị giác máy tính, và việc nghiên cứu về nó có thể mang lại nhiều giá trị cho các dự án ML và doanh nghiệp mà chúng em đang hướng tới.

1. Sơ đồ tổng quan trình bày (phân rã chức năng)

Dự án của nhóm đi qua những phần chính sau :

* Tổng quan
* Lý thuyết thuật toán
* Chuẩn bị dữ liệu
* Kết quả và đánh giá thuật toán
* Kết luận và hướng phát triển

A computer screen shot of a diagram

Description automatically generated

Hình 1. 1 Sơ đồ phân rã chức năng (link xem: [full\_image\_bfd](https://drive.google.com/file/d/1JiWttyQsGsUsrSaR59o7TChAX_2_5iBJ/view?usp=sharing))

1. Công cụ thực hiện
   1. Giới thiệu Python
2. Giới thiệu chung về Python

Python là một ngôn ngữ lập trình mạnh mẽ, đa năng và dễ học. Được phát triển vào cuối những năm 1980 bởi Guido van Rossum, Python nhanh chóng trở thành một trong những ngôn ngữ phổ biến nhất trên thế giới và được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau.

Một trong những ưu điểm nổi bật của Python là cú pháp rõ ràng và dễ đọc, giúp cho việc viết mã trở nên dễ dàng và dễ hiểu. Python cung cấp một loạt các cấu trúc dữ liệu và phương pháp lập trình mạnh mẽ như danh sách (list), từ điển (dictionary), tuple và bộ (set), cho phép lập trình viên làm việc với dữ liệu một cách linh hoạt. Ngoài ra, Python cũng có một hệ sinh thái phong phú các thư viện và frameworks, cho phép lập trình viên xây dựng các ứng dụng phức tạp một cách nhanh chóng và hiệu quả. Ví dụ, thư viện NumPy cung cấp các công cụ mạnh mẽ cho tính toán khoa học và toán học, trong khi thư viện Pandas hỗ trợ xử lý và phân tích dữ liệu. TensorFlow và PyTorch là hai thư viện phổ biến trong lĩnh vực học sâu (deep learning), trong khi Django và Flask là các frameworks phổ biến để phát triển ứng dụng web…

Python có một ưu điểm lớn là tính di động, nghĩa là nó có thể chạy trên nhiều nền tảng khác nhau, bao gồm Windows, macOS và Linux. Điều này giúp cho việc phát triển và triển khai ứng dụng dễ dàng và thuận tiện. Được đánh giá cao với cộng đồng lập trình viên rộng lớn và nhiệt tình. Cộng đồng này cung cấp hàng ngàn tài liệu, ví dụ và hỗ trợ trực tuyến thông qua các diễn đàn, nhóm người dùng và trang web chia sẻ kiến thức.

1. Python trong lĩnh vực học máy

Trong lĩnh vực học máy (ML), Python đã trở thành một ngôn ngữ phổ biến và được ưa chuộng nhờ vào tính linh hoạt, hiệu suất và hệ sinh thái phong phú mà nó cung cấp. Với cú pháp đơn giản và dễ đọc, Python cho phép người dùng tập trung vào việc xây dựng mô hình và thử nghiệm các thuật toán học máy một cách dễ dàng.

Python cung cấp nhiều thư viện và frameworks mạnh mẽ để phát triển các ứng dụng học máy. Một trong số đó là scikit-learn, một thư viện phổ biến được sử dụng để triển khai các thuật toán học máy cơ bản và tiên tiến như K-Nearest Neighbors (KNN), Decision Trees, Random Forests và Support Vector Machines. Scikit-learn cung cấp các công cụ cho việc tiền xử lý dữ liệu, đánh giá mô hình và tinh chỉnh tham số, giúp người dùng dễ dàng xây dựng và đánh giá các mô hình học máy.

Trong số các thuật toán học máy, K-Nearest Neighbors (KNN) là một thuật toán phân loại đơn giản và hiệu quả. Nó dựa trên việc tính toán khoảng cách giữa các điểm dữ liệu để xác định lớp phân loại của điểm dữ liệu mới. Python cung cấp các công cụ mạnh mẽ để thực hiện KNN, và scikit-learn là một trong những thư viện phổ biến cho việc triển khai chúng. Với sự trợ giúp của scikit-learn, việc xây dựng và đánh giá mô hình KNN trở nên đơn giản và thuận tiện. Python không chỉ hỗ trợ trong việc triển khai thuật toán KNN, mà còn là công cụ lý tưởng cho việc khám phá dữ liệu và tiền xử lý. Với thư viện như NumPy và Pandas, người dùng có thể thực hiện các phép toán số học, xử lý dữ liệu và trực quan hóa một cách dễ dàng. Điều này giúp người dùng có thể chuẩn bị dữ liệu cho mô hình KNN và tối ưu hóa quy trình làm việc của họ.

* 1. Các thư viện sử dụng

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thư viện** | **Chức năng chính** |
| 1 | Pandas | Pandas được thiết kế chủ yếu để xử lý và phân tích dữ liệu. Chức năng chính của Pandas là cung cấp cấu trúc dữ liệu linh hoạt như DataFrame và Series để dễ dàng thao tác và phân tích dữ liệu tabular: DataFrame và Series, Trực quan hóa dữ liệu, Thao tác dữ liệu với thời gian, Đánh chỉ số và multilndexing, Xử lý dữ liệu thiếu, Kết hợp và gộp dữ liệu,... |
| 2 | Numpy | NumPy tập trung vào cung cấp đối tượng mảng nhiều chiều và hỗ trợ các phép toán toán học hiệu quả trên các mảng này. Chức năng chính của NumPy là cung cấp một nền tảng mạnh mẽ cho tính toán khoa học và số học: Mảng, Thao tác mảng, Phép toán toán học, Truy cập và thay đổi phần tử mảng, Đại số tuyến tính, Tạo mảng, Thao tác dữ liệu thời gian, Hỗ trợ C-API,.. |
| 3 | Math | Thư viện math trong Python cung cấp các hàm toán học cơ bản. Chức năng chính là hỗ trợ các phép toán và tính toán toán học trong các chương trình Python: Hàm số cơ học, Hàm số làm tròn và số nguyên, Hàm Trigonometry, Hàm số ngẫu nhiên, Hàm số tự nhiên,... |
| 4 | Datetime | Thư viện này cung cấp các lớp và hàm để làm việc với ngày tháng và thời gian trong Python. Lấy thông tin thời gian, Định dạng chuỗi thời gian, Các phép toán thời gian, lớp ‘datetime’ và ‘date’,... |
| 5 | Pyplot | Vẽ đồ thị cơ bản, Định dạng biểu đồ, Định dạng dữ liệu, Biểu đồ phân phối, Lưu và hiển thị biểu đồ,... |
| 6 | Seaborn | cung cấp một giao diện cao cấp để vẽ các biểu đồ thống kê hấp dẫn và dễ đọc: Biểu đồ phân phối, Biểu đồ tương quan, Biểu đồ phân loại, Thiết lập giao diện,... |
| 7 | Optuna | được thiết kế để tìm kiếm không gian siêu tham số để tối ưu hóa mục tiêu cụ thể, thường là một hàm mất mát trong quá trình huấn luyện mô hình máy học: Tìm kiếm tối ưu siêu tham số, Quản lý thông tin và kết quả thử nghiệm, Kiểm tra kết quả tìm kiếm,.... |
| 8 | sklearn.metrics | Cung cấp các hàm và phương thức để đánh giá hiệu suất của mô hình học máy   * confusion\_matrix * classification\_report |
| 9 | sklearn.model\_selection | Cung cấp phương thức train\_test\_split giúp chia dữ liệu thành tập train và tập thử. |
| 10 | Sklearn.neighbours | Cung cấp phương thức KneighborsClassifier giúp  So sánh với mô hình tự xây dựng. |
| 11 | Plotly | Plotly là thư viện Python mạnh mẽ cho việc tạo đồ thị và biểu đồ tương tác. Nó hỗ trợ nhiều loại biểu đồ, tích hợp dễ dàng với các framework web, và có khả năng tương tác trực tiếp trên nền web. Plotly cũng được sử dụng rộng rãi cho việc thực hiện trực tiếp trong Jupyter Notebooks và hỗ trợ nhiều ngôn ngữ lập trình khác nhau. |
| 12 | Streamlit | Streamlit là một framework mã nguồn mở được sử dụng để xây dựng ứng dụng web với Python một cách nhanh chóng và dễ dàng. Đây là một công cụ phổ biến cho việc tạo ra các ứng dụng dựa trên dữ liệu hoặc máy học mà không cần kiến thức sâu về front-end. Bằng cách sử dụng các thư viện như Pandas, Matplotlib và Plotly, nó có thể hiển thị và tương tác với dữ liệu một cách linh hoạt. Điểm mạnh của Streamlit là khả năng tập trung vào logic của ứng dụng. |

Bảng 1. 1 Thư viện sử dụng

* 1. Tổng quan về dữ liệu sử dụng

Dữ liệu chính ở trong bài để xây dựng thuật toán được lấy từ trang web y tế : [*UCI-EDU-DIANOSTIC*](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic) *.*Dữ liệu này mô tả về các đặc điểm được tính toán từ hình ảnh số hóa của chọc hút bằng kim nhỏ (FNA) của khối vú. Chúng mô tả đặc điểm của nhân tế bào có trong hình ảnh.

Bằng dữ liệu này người ta có thể sử dụng để tiến hành chẩn đoán ung thư.

1. Thông tin dự án của nhóm và phân chia nhiệm vụ

Nhóm thực hiện đề tài trong vòng 2 tuần, với 3 thành viên chính, mỗi thành viên đảm nhiệm những nhiệm vụ khác nhau như bảng 1.2. Nhóm đã thực hiện tải dữ liệu và tiến hành phân tích dự báo, tất cả công cụ, thư viện và những yếu tố phần mềm cần thiết để demo dữ liệu nhóm đã tiến hành đóng gói lại vào dự án cộng đồng trên nền tảng GitHub.

1. Giới thiệu về nền tảng Github và dự án của nhóm

GitHub là một nền tảng lưu trữ mã nguồn và quản lý dự án phổ biến dành cho các nhà phát triển phần mềm. Nó cho phép người dùng lưu trữ mã nguồn của dự án, theo dõi sự thay đổi của dữ liệu.

*Các tính năng chính của GitHub bao gồm:*

* Lưu trữ mã nguồn: Người dùng có thể tạo và quản lý các kho lưu trữ (repositories) để lưu trữ mã nguồn của dự án.
* Quản lý phiên bản: GitHub ghi lại lịch sử thay đổi của mã nguồn, cho phép người dùng theo dõi các sửa đổi, trở về phiên bản cũ, và hợp nhất các thay đổi.
* Hợp tác và Phản hồi: Nền tảng này cho phép các nhà phát triển cùng làm việc trên dự án, đưa ra đề xuất (pull request), xem xét và thảo luận về mã nguồn.
* Công cụ Quản lý Dự án: GitHub cung cấp các công cụ quản lý dự án như quản lý nhiệm vụ (issues), wiki, và dự án kanban board để theo dõi tiến độ công việc.

GitHub đã trở thành một trong những công cụ quan trọng cho cộng đồng phát triển phần mềm, cung cấp nền tảng cho sự hợp tác, kiểm soát phiên bản, và lưu trữ mã nguồn dự án một cách dễ dàng và hiệu quả.y đổi, hợp tác với đồng nghiệp và kiểm soát phiên bản.

*Dự án của nhóm trên nền tảng GitHub:*

Link truy cập dự án:[*GITHUB-DIANOSING CANCER WITH KNN*](https://github.com/davisduccopny/Diagnosing-cancer-with-the-KNN-algorithm/tree/main)

* Thông tin cách cài đặt và sử dụng: README.md
* Thông tin các thư viện cần thiết: requirements.txt
* Cấu trúc thư mục dự án github:

A black background with white text

Description automatically generated

Hình 1. 2 Cấu trúc thư mục dự án Github

1. Phân chia nhiệm vụ

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Nhiệm vụ** |
| 1 | Hoàng Xuân Quốc | 2156210125 | Phân tích mô tả, code lại mô hình thuật toán |
| 2 | Nguyễn Viết Đức | 2156210100 | Chuẩn bị dữ liệu, code lại mô hình thuật toán |
| 3 | Đặng Hoàng Chiến | 2156210095 | Làm báo cáo, code mô hình thuật toán |

Bảng 1. 2 Phân công nhiệm vụ

1. LÝ THUYẾT THUẬT TOÁN
2. Giới thiệu về data mining và bài toán phân loại/phân cụm/tìm luật kết hợp
3. Thuật toán K – nearest-neibourgh
   1. Giới thiệu về thuật toán và bài toán mà thuật toán giải quyết
      1. Tổng quan về thuật toán
4. Khái niệm

Thuật toán K-Nearest Neighbor là thuật toán có nghĩa tiếng Việt là K láng giềng gần nhất, viết tắt là KNN. Thuật toán KNN là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có. Thuật toán này được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau từ phân loại, hồi quy cho đến hệ thống gợi ý.

1. Nguyên lý hoạt động cơ bản

KNN hoạt động dựa trên nguyên tắc rằng các điểm dữ liệu tương tự thường nằm gần nhau. Để dự đoán nhãn cho một điểm dữ liệu mới, thuật toán này xem xét ‘K’ điểm dữ liệu gần nhất và dựa trên đa số nhãn của chúng để xác định nhãn cho điểm dữ liệu mới đó.

Giá trị ‘K’ là một siêu tham số quan trọng, quy định số lượng hàng xóm gần nhất mà thuật toán sẽ xem xét. Việc chọn K đúng là rất quan trọng: K quá nhỏ có thể làm mô hình bị ảnh hưởng bởi nhiễu dữ liệu, trong khi K quá lớn có thể làm mô hình không nhạy với đặc điểm cụ thể của dữ liệu.

Tính toán khoảng cách giữa các điểm dữ liệu là một phần quan trọng của KNN dùng để đo lường sự tương tự giữa các điểm dữ liệu. Phương pháp phổ biến nhất là khoảng cách Euclidean. Tuy nhiên, có nhiều phương pháp khác như khoảng cách Manhattan được sử dụng tùy thuộc vào bản chất của dữ liệu.

1. Lịch sử hình thành và phát triển

Khởi nguyên: Nguồn gốc và lịch sử phát triển của thuật toán K-Nearest Neighbors (KNN) bắt đầu trong giai đoạn đầu của lĩnh vực học máy và trí tuệ nhân tạo. KNN xuất hiện lần đầu vào những năm 1950 và 1960 trong quá trình nghiên cứu về mô hình hóa mẫu và nhận dạng. Mặc dù không rõ người chính thức sáng tạo KNN, thuật toán này nhanh chóng trở nên phổ biến do tính đơn giản và dễ hiểu của nó.

Phát triển và cải tiến: Trong thập kỷ tiếp theo, KNN đã được cải tiến và tinh chỉnh để nâng cao hiệu suất và độ chính xác. Các nghiên cứu tập trung vào việc tối ưu hóa lựa chọn giá trị 'K', phương pháp tính khoảng cách và xử lý dữ liệu lớn. Một trong những cải tiến quan trọng là áp dụng các kỹ thuật giảm chiều dữ liệu như Principal Component Analysis (PCA) và t-SNE để giảm tải tính toán và cải thiện khả năng phân loại.

* + 1. Vấn đề mà bài toán giải quyết ( về mặt lý thuyết và thực tế)

1. Về mặt lý thuyết

Về mặt lý thuyết, bài toán mà thuật toán K-Nearest Neighbors giải quyết là bài toán phân loại và gợi ý dựa trên các đặc trưng của dữ liệu. KNN được sử dụng để xác định lớp hoặc nhãn của một điểm dữ liệu mới dựa trên các điểm dữ liệu đã biết trước trong tập huấn luyện. KNN thuộc vào lớp các thuật toán học máy không giám sát (unsupervised learning) và phụ thuộc vào việc tính toán khoảng cách giữa các điểm dữ liệu. Thuật toán này dựa trên giả định rằng các điểm dữ liệu cùng lớp sẽ có tính chất tương tự và nằm gần nhau trong không gian đặc trưng.

1. Về những ứng dụng thực tế

* Phân loại văn bản: sử dụng để phân loại văn bản, chẳng hạn như phân loại email vào hộp thư rác hoặc không phải hộp thư rác dựa trên nội dung và các đặc trưng của email.
* Nhận dạng khuôn mặt: KNN có thể được áp dụng trong hệ thống nhận dạng khuôn mặt để xác định xem một khuôn mặt mới thuộc về người nào dựa trên các khuôn mặt đã biết trước trong tập huấn luyện.
* Hệ thống gợi ý: Có thể được sử dụng trong các hệ thống gợi ý sản phẩm hoặc nội dung. Điều này có thể bao gồm việc đề xuất các sản phẩm tương tự cho người dùng dựa trên lịch sử mua hàng hoặc đánh giá của họ.
* Phát hiện gian lận tín dụng: KNN có thể được sử dụng để phát hiện gian lận tín dụng bằng cách xác định xem một giao dịch mới có khả năng là gian lận dựa trên các đặc trưng của giao dịch và thông tin từ các giao dịch trước đó.
* Phân loại hình ảnh: Được áp dụng trong việc phân loại hình ảnh vào các nhóm khác nhau. Ví dụ, xác định xem một bức ảnh là chó, mèo hoặc chim cút dựa trên các đặc trưng hình ảnh và tập huấn luyện các hình ảnh đã được gán nhãn trước đó.
* Nhận dạng tiếng nói: áp dụng trong các hệ thống nhận dạng tiếng nói để xác định người nói dựa trên các đặc trưng âm thanh của giọng nói.
* Phân loại sản phẩm: KNN có thể được sử dụng trong các cửa hàng trực tuyến để phân loại sản phẩm vào các danh mục khác nhau dựa trên các đặc trưng của sản phẩm như mô tả, giá, nhãn hiệu, đánh giá và thông tin liên quan khác.
* Dự báo thời tiết: nó có thể được áp dụng trong việc dự báo thời tiết dựa trên các đặc trưng thời tiết trước đó như nhiệt độ, độ ẩm, áp suất không khí và các yếu tố khí hậu khác.
* Hỗ trợ quyết định y tế: KNN có thể được sử dụng để hỗ trợ quyết định trong lĩnh vực y tế, chẳng hạn như xác định liệu một bệnh nhân có nguy cơ cao mắc bệnh tim mạch hay không dựa trên các chỉ số sức khỏe và lịch sử bệnh lý…
  1. Đặc trưng của thuật toán ( ưu điểm và nhược điểm đối với từng trường hợp cụ thể)

1. Đặc trưng chung

Dựa trên sự tương tự: KNN sẽ đo lường tính toán độ tương đồng hoặc khoảng cách giữa các điểm dữ liệu trong không gian đặc trưng. Nó sẽ sử dụng một phép đo khoảng cách để xác định các điểm gần nhất với điểm dữ liệu đang được dự đoán.Các điểm gần nhất này sau đó được sử dụng để suy ra nhãn hoặc giá trị dự đoán cho điểm dữ liệu đó.

* Ưu điểm: Với đặc trưng này, việc xác định và tính toán giữa các điểm dữ liệu là khá đơn giản và dễ hiểu. Nó cho phép đo lường sự tương đồng dựa trên khoảng cách giữa các điểm, giúp xác định các điểm dữ liệu gần nhau.
* Nhược điểm: Nó có thể không hoạt động tốt đối với các tập dữ liệu có đặc trưng không đồng nhất hoặc khi các lớp không phân tách rõ ràng trong không gian đặc trưng của dữ liệu.

Không giám sát: KNN là một thuật toán học không giám sát, nghĩa là nó không yêu cầu một giai đoạn huấn luyện. Nó chỉ cần dữ liệu huấn luyện đã được gán nhãn để thực hiện dự đoán.

* Ưu điểm: Chúng ta chỉ cần dữ liệu huấn luyện đã được gán nhãn để thực hiện dự đoán, giảm thiểu được kha khá công việc chuẩn bị dữ liệu huấn luyện.
* Nhược điểm: Vì KNN không yêu cầu huấn luyện tường minh, nó có thể không học được các mẫu, quy luật hoặc đặc trưng phức tạp trong dữ liệu. Điều này có thể dẫn đến hiệu suất dự đoán không tốt trên các tập dữ liệu phức tạp.

Khoảng cách: KNN dựa trên các điểm dữ liệu gần nhất trong không gian đặc trưng để đưa ra dự đoán. Giá trị của K trong thuật toán KNN xác định số lượng các hàng xóm gần nhất cần xem xét.

* Ưu điểm: KNN dựa trên các điểm dữ liệu gần nhất để đưa ra dự đoán. Điều này có nghĩa là KNN tập trung vào các điểm dữ liệu có tương đồng về mặt không gian đặc trưng, giúp xác định nhóm và phân loại dữ liệu.
* Nhược điểm: Việc tìm kiếm các hàng xóm gần nhất trong quá trình dự đoán có độ phức tạp tính toán cao. Đặc biệt là khi số lượng điểm dữ liệu huấn luyện lớn, việc tìm kiếm các hàng xóm xung quanh có thể tốn nhiều thời gian và tài nguyên.

Linh hoạt với số chiều: KNN có khả năng làm việc tốt với các tập dữ liệu có số chiều cao. Điều này là do KNN không giả định về phân phối của dữ liệu và không áp dụng một mô hình toán học phức tạp. Tuy nhiên, khi số chiều tăng lên, việc tính toán khoảng cách có thể trở nên tốn kém và hiệu quả của thuật toán có thể giảm. Để giải quyết ta có thể sử dụng kỹ thuật giảm chiều dữ liệu như PCA hoặc LLE trước khi áp dụng KNN. Điều này giúp giảm số chiều của dữ liệu trong khi vẫn giữ lại thông tin quan trọng và cấu trúc của dữ liệu.

* Ưu điểm: KNN có khả năng làm việc tốt với các tập dữ liệu có số chiều cao. Điều này là do KNN không giả định về phân phối của dữ liệu và không áp dụng một mô hình toán học phức tạp. Nó có thể áp dụng trực tiếp trên không gian đặc trưng có số chiều cao mà không cần quá nhiều điều chỉnh.
* Nhược điểm: Mặc dù linh hoạt với số chiều, KNN có thể gặp khó khăn trong việc xử lý các tập dữ liệu có số chiều rất lớn. Việc tính toán khoảng cách giữa các điểm dữ liệu có thể trở nên rất tốn kém và không hiệu quả trong trường hợp này.

1. Ưu điểm chung

* Đơn giản và dễ hiểu: KNN là một thuật toán đơn giản và dễ hiểu. Nguyên tắc hoạt động của nó dựa trên ý tưởng rằng các điểm dữ liệu gần nhau trong không gian đặc trưng có khả năng thuộc cùng một nhãn.
* Khả năng xử lý dữ liệu phi cấu trúc: KNN có khả năng xử lý dữ liệu phi cấu trúc, không yêu cầu các giả định về cấu trúc dữ liệu. Điều này cho phép nó áp dụng cho nhiều loại dữ liệu, bao gồm cả dữ liệu văn bản, hình ảnh, và âm thanh.
* Khả năng mở rộng: Thuật toán này không yêu cầu huấn luyện tường minh, do đó nó có khả năng mở rộng cho việc thêm dữ liệu mới một cách dễ dàng.
* Hiệu suất tốt đối với tập dữ liệu nhỏ: KNN hoạt động tốt trên các tập dữ liệu nhỏ với số lượng mẫu ít. Nó không đòi hỏi quá nhiều tính toán trước khi thực hiện dự đoán.
* Dễ dàng tinh chỉnh tham số: K là tham số quan trọng trong thuật toán KNN. Việc điều chỉnh giá trị K có thể ảnh hưởng đến hiệu suất của thuật toán. Tuy nhiên, việc tinh chỉnh K là khá dễ dàng và có thể được thực hiện thông qua quá trình thử và sai
* Sử dụng được trong cả phân loại và hồi quy: KNN có thể được sử dụng cho cả hai mục đích là phân loại và hồi quy.
* Hoạt động tốt trong trường hợp phân loại với nhiều lớp: KNN có thể được sử dụng để phân loại dữ liệu với nhiều lớp.
* Không cần giả định gì về phân phối của các class: KNN là một thuật toán phi tham số, vì vậy nó không cần giả định gì về phân phối của các class. Điều này giúp KNN có thể hoạt động tốt với các tập dữ liệu có phân phối phức tạp hoặc không xác định

1. Nhược điểm chung

* Phụ thuộc vào kích thước dữ liệu: hiệu suất sẽ giảm khi tập dữ liệu lớn vì việc tính toán khoảng cách giữa các điểm dữ liệu mất nhiều thời gian. KNN không phải là thuật toán phù hợp cho việc xử lý các tập dữ liệu lớn.
* Tính toán tài nguyên: KNN có đòi hỏi khá nhiều tài nguyên tính toán khi tập dữ liệu lớn, đặc biệt là khi số chiều của không gian đặc trưng cao.
* Nhạy cảm với nhiễu và dữ liệu không đồng nhất: Dễ bị ảnh hưởng bởi các nhiễu trong dữ liệu và các điểm dữ liệu nằm trong các lớp khác nhau gần nhau. Điều này có thể dẫn đến việc phân loại không chính xác hoặc không ổn định.
* Cần xử lý các biến số và mất cân bằng dữ liệu: KNN không xử lý được các biến số khác nhau và cần sự cân bằng dữ liệu trong các lớp khác nhau. Nếu một lớp có số lượng mẫu nhiều hơn so với lớp khác, KNN có thể dễ dàng bị thiên vị và cho ra kết quả không chính xác.
* Lưu trữ dữ liệu: KNN cần lưu trữ toàn bộ tập dữ liệu huấn luyện trong bộ nhớ để thực hiện việc tìm kiếm hàng xóm gần nhất. Điều này có thể là một vấn đề khi tập dữ liệu rất lớn.
* Lựa chọn tham số K: Tham số K trong KNN cần được chọn một cách thích hợp. Một K quá nhỏ có thể dẫn đến overfitting, trong khi một K quá lớn có thể dẫn đến underfitting.
  1. Nguyên lý thuật toán ( trình bày sơ đồ, công thức và giải thích từng phần)

Nguyên lý hoạt động của thuật toán K-Nearest Neighbors (KNN) là dựa trên giả định rằng các điểm dữ liệu có tính chất tương đồng với nhau thường nằm gần nhau trong không gian đặc trưng. Khi có một điểm dữ liệu mới, KNN sẽ tìm ra các điểm dữ liệu gần nó nhất trong không gian đặc trưng, và sử dụng nhãn của các điểm dữ liệu này để dự đoán nhãn của điểm dữ liệu mới.

A yellow diagram with black text

Description automatically generated

Hình 1. 3 Sơ đồ nguyên lý

***Giải thích chi tiết sơ đồ:***

* Chuẩn bị dữ liệu huấn luyện: Thuật toán KNN yêu cầu một tập dữ liệu huấn luyện đã được gán nhãn, trong đó mỗi điểm dữ liệu có các đặc trưng và nhãn tương ứng
* Xác định giá trị 'K': Người dùng cần quyết định giá trị 'K', tức là số lượng hàng xóm gần nhất mà thuật toán sẽ xem xét trong quá trình dự đoán.
* Tính toán khoảng cách: KNN sử dụng một phương pháp tính khoảng cách để đo lường sự tương tự giữa các điểm dữ liệu. Khoảng cách được tính giữa điểm dữ liệu mới và các điểm dữ liệu trong tập huấn luyện.
* Các phép đo khoảng cách chủ yếu trong KNN là
* Khoảng cách Euclidean: Đây là phép đo khoảng cách phổ biến nhất trong KNN. Khoảng cách Euclidean giữa hai điểm dữ liệu trong không gian đặc trưng được tính bằng:

A math formula with red dots

Description automatically generated

  Trong đó, Ai​ và Bi​ là giá trị của đặc trưng tương ứng ở vị trí thứ i, và n là số lượng đặc trưng

* Khoảng cách Manhattan: Còn được gọi là khoảng cách thành phố, khoảng cách Manhattan tính toán tổng các độ lệch tuyệt đối giữa các giá trị của các đặc trưng tương ứng của hai điểm dữ liệu với công thức.

A white text on a black background

Description automatically generated

Trong đó, Ai​ và Bi​ là giá trị của đặc trưng tương ứng ở vị trí thứ i, và n là số lượng đặc trưng

* Khoảng cách Mahalanobis: Khoảng cách Mahalanobis tính toán khoảng cách giữa hai điểm dữ liệu dựa trên sự khác biệt giữa chúng và ma trận hiệp phương sai của dữ liệu. Đây là một phép đo mạnh mẽ khi dữ liệu có các đặc trưng không đồng nhất hoặc có tương quan. Nó được thể hiện bằng công thức:

A white text on a blue background

Description automatically generated

Trong đó, A và B là vectơ đặc trưng của hai điểm dữ liệu, S là ma trận hiệp phương sai của dữ liệu, và T biểu thị phép chuyển vị (transpose) của vector.

* Khoảng cách Cosine: Khoảng cách Cosine đo lường góc giữa hai vectơ đặc trưng và được sử dụng để đo lường độ tương đồng giữa các điểm dữ liệu. Nó thường được sử dụng trong các tập dữ liệu văn bản hoặc trong trường hợp đặc trưng là các giá trị đại diện cho hướng. Nó được thể hiện bằng công thức:

A close-up of a logo

Description automatically generated

Trong đó, *A*⋅*B* là tích vô hướng của hai vectơ,∥*A*∥ và ∥*B*∥ là độ dài của các vectơ tương ứng. Công thức này thể hiện mức độ tương đồng giữa hai vectơ, dựa trên góc giữa chúng trong không gian đặc trưng. Khoảng cách Cosine thường cho kết quả trong khoảng từ 0 đến 1, với 0 thể hiện hai vectơ không tương đồng và 1 thể hiện chúng hoàn toàn giống nhau.

* Xác định 'K' hàng xóm gần nhất: KNN chọn 'K' điểm dữ liệu gần nhất (hàng xóm gần nhất) dựa trên khoảng cách tính được.
* Đa số phiếu bầu: KNN sử dụng đa số phiếu bầu để quyết định nhãn cho điểm dữ liệu mới. Điểm dữ liệu mới sẽ được gán nhãn bằng nhãn xuất hiện nhiều nhất trong 'K' hàng xóm gần nhất.
* Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng cách đo lường độ chính xác, độ phủ (recall), độ chính xác (precision) và các độ đo khác trên tập dữ liệu kiểm tra hoặc sử dụng các kỹ thuật cross-validation
* Tinh chỉnh tham số: Nếu kết quả đánh giá không đạt yêu cầu, bạn có thể tinh chỉnh tham số như K, phương pháp đo khoảng cách hoặc tiền xử lý dữ liệu để cải thiện hiệu suất của mô hình.
  1. Minh họa thuật toán
     1. Dữ liệu

|  |  |  |
| --- | --- | --- |
| **Tuổi** | **Thu nhập** | **Đối tượng khách hàng** |
| 15 | 30 Triệu | Khách hàng không tiềm năng |
| 20 | 100 Triệu | Khách hàng tiềm năng |
| 22 | 70 Triệu | Khách hàng không tiềm năng |
| 24 | 120 Triệu | Khách hàng tiềm năng |
| 30 | 200 Triệu | Khách hàng tiềm năng |
| 35 | 110 Triệu | Khách hàng không tiềm năng |
| 40 | 300 Triệu | Khách hàng tiềm năng |
| 45 | 90 Triệu | Khách hàng không tiềm năng |
| 50 | 400 Triệu | Khách hàng tiềm năng |
| 55 | 420 Triệu | Khách hàng không tiềm năng |
| 27 | 80 Triệu | ??? |

Bảng 2. 1 Dữ liệu demo

Dữ liệu cần phần loại là ( 27 tuổi, thu nhập 80 Triệu ). Chúng ta dự đoán xem khách hàng này sẽ là khách hàng tiềm năng hay là khách hàng không tiềm năng.

* + 1. Các bước thực hiện

1. Với độ đo Euclidean

**Bước 1:** Chuẩn bị dữ liệu

**Bước 2:** Chọn tham số K

Trong ví dụ này chúng ta sẽ chọn tham số K=3

**Bước 3:** Tính khoảng cách giữa dữ liệu mới với tất cả các dữ liệu còn lại: ở đây chúng ta sẽ sử dụng công thức Euclidean để tính toán khoảng cách:

A square and square equation

Description automatically generated

Euclidean = sqrt (27-15)^2+(80-30)^2=51.4…

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuổi** | **Thu nhập** | **Đối tượng khách hàng** | **Khoảnh cách** |
| 15 | 30 Triệu | Khách hàng không tiềm năng | 51.4 |
| 20 | 100 Triệu | Khách hàng tiềm năng | 21.2 |
| 22 | 70 Triệu | Khách hàng không tiềm năng | 11.2 |
| 24 | 120 Triệu | Khách hàng tiềm năng | 40.1 |
| 30 | 200 Triệu | Khách hàng tiềm năng | 120 |
| 35 | 110 Triệu | Khách hàng không tiềm năng | 31.1 |
| 40 | 300 Triệu | Khách hàng tiềm năng | 220.4 |
| 45 | 90 Triệu | Khách hàng không tiềm năng | 20.6 |
| 50 | 400 Triệu | Khách hàng tiềm năng | 320.8 |
| 55 | 420 Triệu | Khách hàng tiềm năng | 341.1 |
| 27 | 80 Triệu | ??? |  |

Bảng 2. 2 Euclidean demo

**Bước 4:** Lựa chọn k điểm có khoảng cách gần nhất: Ở đây chúng ta thấy 3 điểm có khoảng cách gần nhất là:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuổi** | **Thu nhập** | **Đối tượng khách hàng** | **Khoảnh cách** |
| 20 | 100 Triệu | Khách hàng tiềm năng | 21.2 |
| 22 | 70 Triệu | Khách hàng không tiềm năng | 11.2 |
| 45 | 90 Triệu | Khách hàng không tiềm năng | 20.6 |

Bảng 2. 3 Phân tập khách hàng

Với 3 khoảng cách này chúng ta nhận ra rằng 3 kết quả là 2 (Khách hàng không tiềm năng) và 1(Khách hàng tiềm năng).

**Bước 5**: Dự đoán dữ liệu mới dựa trên K gần nhất

Trong 3 kết quả này Khách hàng không tiềm năng xuất hiện nhiều hơn nên chúng ta đưa ra dự đoán khách hàng này là khách hàng không tiềm năng.

1. Với độ đo Manhattan

**Bước 1:** Chuẩn bị dữ liệu

**Bước 2:** Chọn tham số K

Trong ví dụ này chúng ta sẽ chọn tham số K=3

**Bước 3:** Tính khoảng cách giữa dữ liệu mới với tất cả các dữ liệu còn lại: ở đây chúng ta sẽ sử dụng công thức Manhattan để tính toán khoảng cách:

A close-up of a mathematical equation

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuổi** | **Thu nhập** | **Đối tượng khách hàng** | **Khoảnh cách** |
| 15 | 30 Triệu | Khách hàng không tiềm năng | 62 |
| 20 | 100 Triệu | Khách hàng tiềm năng | 27 |
| 22 | 70 Triệu | Khách hàng không tiềm năng | 15 |
| 24 | 120 Triệu | Khách hàng tiềm năng | 43 |
| 30 | 200 Triệu | Khách hàng tiềm năng | 123 |
| 35 | 110 Triệu | Khách hàng không tiềm năng | 38 |
| 40 | 300 Triệu | Khách hàng tiềm năng | 233 |
| 45 | 90 Triệu | Khách hàng không tiềm năng | 28 |
| 50 | 400 Triệu | Khách hàng tiềm năng | 343 |
| 55 | 420 Triệu | Khách hàng tiềm năng | 368 |

Bảng 2. 4 Manhattan demo

**Bước 4:** Lựa chọn k điểm có khoảng cách gần nhất: Ở đây chúng ta thấy 3 điểm có khoảng cách gần nhất là:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuổi** | **Thu nhập** | **Đối tượng khách hàng** | **Khoảnh cách** |
| 20 | 100 Triệu | Khách hàng tiềm năng | 27 |
| 22 | 70 Triệu | Khách hàng không tiềm năng | 15 |
| 45 | 90 Triệu | Khách hàng không tiềm năng | 28 |

Bảng 2. 5 Phân tập khách hàng manhattan

Với 3 khoảng cách này chúng ta nhận ra rằng 3 kết quả là 2 (Khách hàng không tiềm năng) và 1(Khách hàng tiềm năng).

**Bước 5:** Dự đoán dữ liệu mới dựa trên K gần nhất

Trong 3 kết quả này Khách hàng không tiềm năng xuất hiện nhiều hơn nên chúng ta đưa ra dự đoán khách hàng này là khách hàng không tiềm năng.

1. Với độ đo Minkowski

A green square with black and white text

Description automatically generated

***Trong ví dụ này:***

p=1 đối với công thức Manhattan kết quả khoảng cách giống với ví dụ đã thực hiện ở trên.

p=2 đối với công thức Euclidean kết quả khoảng cách giống với ví dụ đã thực hiện ở trên.

* + 1. Kết quả
* Với Euclidean: Kết quả dự đoán của thuật toán KNN cho dữ liệu mới là khách hàng không tiềm năng.
* Với Manhattan: Kết quả dự đoán của thuật toán KNN cho dữ liệu mới là khách hàng không tiềm năng.
* Với Minkowski : Kết quả dự đoán của thuật toán KNN cho dữ liệu mới là khách hàng không tiềm năng.

1. CHUẨN BỊ DỮ LIỆU
2. Giới thiệu dữ liệu

* Insert các thư viện được sử dụng trong bài làm.

A screenshot of a computer

Description automatically generated

Hình 3. 1 Insert thư viện sử dụng

* Tải dataset.



Hình 3. 2 Tải dataset

* Mô tả từng cột dữ liệu trong dataset

A screenshot of a computer

Description automatically generated

Hình 3. 3 Mô tả dữ liệu

* Xem 10 dòng đầu tiên của dataset

A screenshot of a computer

Description automatically generated

Hình 3. 4 10 dòng đầu của dữ liệu

1. Tiền xử lý dữ liệu
   1. Làm sạch dữ liệu

* Kiểm tra các dữ liệu thiếu và dữ liệu trống.

A black and grey background

Description automatically generated

Hình 3. 5 Kiểm tra dữ liệu trống

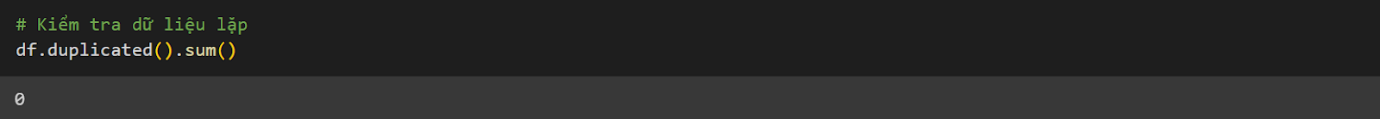
* Xóa dữ liệu trống.

A screenshot of a computer

Description automatically generated

Hình 3. 6 Xóa dữ liệu trống

* Kiểm tra dữ liệu lặp



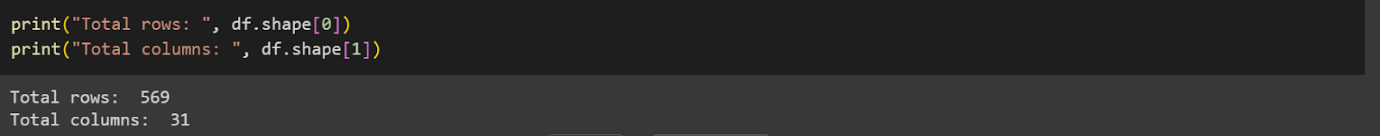
Hình 3. 7 Kiểm tra dữ liệu lặp

* 1. Biến đổi dữ liệu
* Đặt lại chỉ số của DataFrame và loại bỏ chỉ số cũ và loại bỏ cột có tên 'id' khỏi DataFrame và thực hiện thay đổi trực tiếp trên DataFrame hiện tại.



Hình 3. 8 Đặt lại index

* In ra hình dữ liệu:



Hình 3. 9 Đặt hình dạng của dữ liệu

* Xem lại dataset

A screenshot of a computer

Description automatically generated

Hình 3. 10 Xem lại dữ liệu

* Chia dữ liệu để thử nghiệm



Hình 3. 11 Chia dữ liệu thử nghiệm

* Encode tập y



Hình 3. 12 Encode tập y

1. Mô tả dữ liệu (thống kê mô tả)

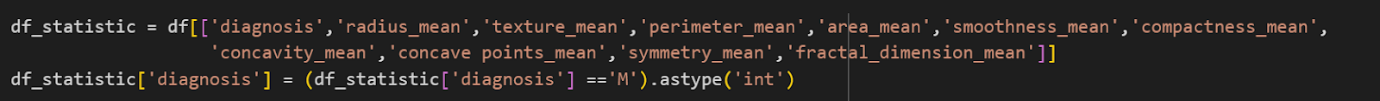
* Đặt ra các câu hỏi thống kê nhằm hiểu rõ hơn về dữ liệu như:
* Count (Số lượng): Số lượng giá trị không rỗng (non-null) trong mỗi cột.
* Mean (Trung bình): Giá trị trung bình của dữ liệu.
* Std (Độ lệch chuẩn): Độ lệch chuẩn của dữ liệu, đo độ biến động của dữ liệu.
* Min (Giá trị nhỏ nhất): Giá trị nhỏ nhất trong dữ liệu.
* Phân vị 25%, 50%, 75%: Các phân vị của dữ liệu, cung cấp thông tin về phân phối của dữ liệu.
* Max (Giá trị lớn nhất): Giá trị lớn nhất trong dữ liệu.

A screenshot of a computer

Description automatically generated

Hình 3. 13 Một số thông tin cơ bản

* ***Giải thích các biến***
  + radius\_mean, texture\_mean, perimeter\_mean, area\_mean : các biến liên quan chặt chẽ tới kích thước và hình dạng của tế bào.
  + smoothness\_mean,compactness\_mean: độ mịn và độ đặc của tế bào , phản ánh cấu trúc tế bào.
  + concavity\_mean, concave points\_mean: Liên quan đến sự uốn cong và các điểm lõm của tế bào.
  + Symmetry\_mean, fractal\_dimension\_mean: tính đối xứng và chiều mịn của tế bào
  + Các biến hậu tố \_se là các giá trị lỗi chuẩn của các đặc trưng trên.
  + Các biến với hậu tố \_worst là các giá trị lớn nhất trong 3 lần đo của các đặc trưng trên.
* Một số quan sát rút ra:
  + Tổng số quan sát là 569.
  + Giá trị cao nhất, thấp nhất của các cột dữ liệu khác nhau đáng kể cho thấy phạm vi giá trị lớn của các biến.
  + Độ lệch chuẩn của từng biến khá lớn, cho thấy mức độ biến động cao của dữ liệu.
* Kiểm tra phân phối dữ liệu



A screenshot of a computer

Description automatically generated

A group of blue bars

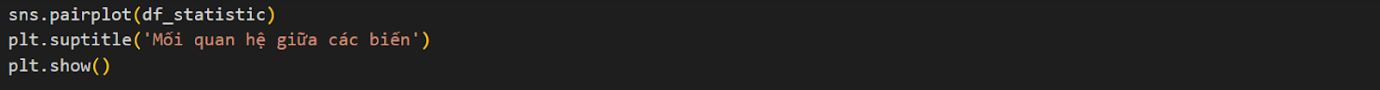
Description automatically generated

A graph of different sizes and colors

Description automatically generated with medium confidence

Hình 3. 14 Kiểm tra phân phối dữ liệu

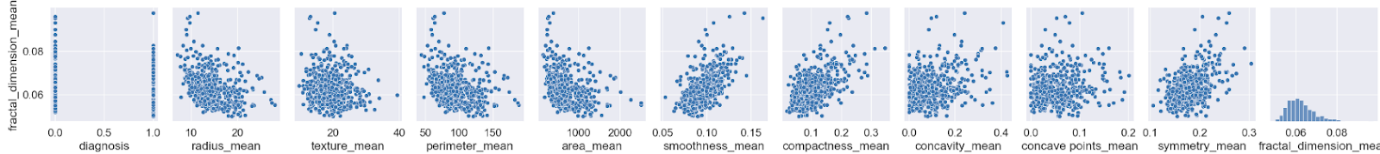
* Độ tương quan giữa các biến:



A collage of blue dots

Description automatically generatedA collage of blue dots

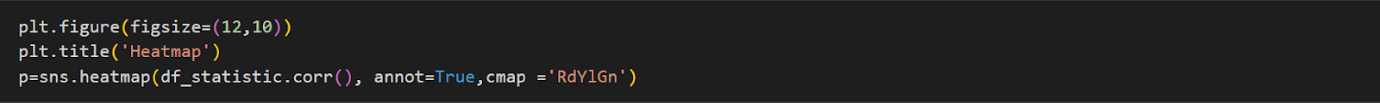
Description automatically generated



Hình 3. 15 Độ tương quan giữa các biến

Hệ số Tương quan Pearson: giúp tìm hiểu mối quan hệ giữa hai lượng. Nó cung cấp đo lường về sức mạnh của mối liên kết giữa hai biến. Giá trị của Hệ số Tương quan Pearson có thể nằm trong khoảng từ -1 đến +1. Giá trị 1 có nghĩa là chúng có mối tương quan cao và giá trị 0 có nghĩa là không có tương quan.

* Vẽ biểu đồ Heatmap



A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Hình 3. 16 Heatmap

1. KẾT QUẢ VÀ ĐÁNH GIÁ THUẬT TOÁN
2. Xây dựng thuật toán
3. Áp dụng thuật toán cho bài toán
4. Thử nghiệm thuật toán
   1. Thử nghiệm với sự thay đổi của k
   2. Thử nghiệm với sự thay đổi của p
5. Đánh giá và so sánh thuật toán
6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN
7. Kết luận
8. Hướng phát triển
9. TÀI LIỆU THAM KHẢO