TRƯỜNG ĐẠI HỌC HỌC VĂN LANG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC HK242

NHẬP MÔN PHÂN TÍCH DỮ LIỆU LỚN (71ITDS40403)

ỨNG DỤNG CÔNG NGHỆ DỮ LIỆU LỚN TRONG VIỆC XỬ LÝ CÔNG VIỆC TRONG CÔNG TY

Nhóm sinh viên thực hiện (Họ tên - Mã SV):

1.Trần Tấn Phát_2274802010644 (Trưởng nhóm)

2. Đăng Võ Quang Huy_2274802010301

3. Huỳnh Gia Huy_2274802010303

4. Bùi Nguyên Tín_2274802010894

5. Nguyễn Thái Nguyên_2274802010587

TP. Hồ Chí Minh – năm 2025

MỤC LỤC

MỞ ĐẦU

1. Lý do chọn chủ đề:

- Trong thời đại công nghiệp 4.0, công nghệ dữ liệu lớn (Big Data) đã trở thành một trong những trụ cột quan trọng thúc đẩy sự phát triển của các doanh nghiệp trên toàn cầu. Đặc biệt, trong lĩnh vực quản trị doanh nghiệp và tối ưu hóa hiệu suất làm việc của nhân viên, việc ứng dụng công nghệ dữ liệu lớn mang lại khả năng nâng cao hiệu quả quản lý và hỗ trợ ra quyết định chiến lược dựa trên dữ liệu thực tế. Các công ty hiện nay phải đối mặt với khối lượng thông tin khổng lồ từ hồ sơ nhân viên, chẳng hạn như thông tin cá nhân (họ tên, ngày sinh, địa chỉ), dữ liệu công việc (chức danh, phòng ban, lương), và các chỉ số hiệu suất (điểm hiệu suất, số ngày nghỉ) như được thể hiện trong tệp CSV.
- Việc xử lý thủ công các dữ liệu này – ví dụ, thông tin của hơn 200 nhân viên với các thuộc tính đa dạng từ Kathy May (Social Worker, Finance) đến Timothy Burke (Insurance Underwriter, Finance) – không chỉ tốn thời gian mà còn dễ dẫn đến sai sót. Điều này tạo ra nhu cầu cấp thiết phải ứng dụng công nghệ dữ liệu lớn để tự động hóa, phân tích và tối ưu hóa quy trình xử lý công việc.
- Chủ đề "Ứng dụng công nghệ dữ liệu lớn trong việc xử lý công việc của nhân viên trong công ty" được chọn vì tính thực tiễn cao, đáp ứng nhu cầu thực tế của doanh nghiệp trong việc nâng cao năng suất lao động, giảm thiểu lãng phí tài nguyên, và cải thiện trải nghiệm làm việc của nhân viên.
- Với dữ liệu từ tệp CSV, việc phân tích các chỉ số như Performance Score, Days Off, và Status có thể giúp doanh nghiệp hiểu rõ hơn về hiệu suất và tình trạng nhân viên, từ đó đưa ra các điều chỉnh kịp thời. Đây cũng là một lĩnh vực đang phát triển mạnh mẽ, phù hợp với xu hướng số hóa và chuyển đổi kỹ thuật số mà các công ty hiện đại, bao gồm cả những công ty nổi tiếng đang hướng tới.

2. Đối tượng, phạm vi tìm hiểu

- Đối tượng nghiên cứu của chủ đề này là các công ty vừa và lớn, nơi có số lượng nhân viên đáng kể và khối lượng dữ liệu công việc cần xử lý, tương tự như các công ty được liệt kê trong tệp CSV. Dữ liệu từ tệp bao gồm thông tin của nhân viên từ nhiều phòng ban như Finance, IT, HR, Marketing, Operations, và Sales, phản ánh thực tế của một doanh nghiệp điển hình. Phạm vi tìm hiểu tập trung vào việc ứng dụng công nghệ dữ liệu lớn trong các khía cạnh như:

    - Quản lý hiệu suất nhân viên (dựa trên Performance Score).
    - Phân bổ công việc (dựa trên Status và Days Off).
    - Dự đoán nhu cầu nhân sự (phân tích xu hướng nghỉ việc từ Status: Active, On Leave, Resigned).
    - Cải thiện quy trình nội bộ (sử dụng dữ liệu từ Department và Salary để tối ưu hóa nguồn lực).

- Về lịch sử và quá trình phát triển, công nghệ dữ liệu lớn bắt nguồn từ những năm 2000 khi các tập đoàn như Google và Amazon tiên phong trong việc khai thác dữ liệu người dùng để tối ưu hóa dịch vụ. Đến nay, Big Data đã mở rộng ứng dụng sang quản trị nhân sự và tổ chức công việc, như việc xử lý thông tin nhân viên tương tự dữ liệu trong tệp CSV. Sự phát triển của các công cụ như Hadoop, Spark, và các nền tảng đám mây đã giúp các công ty vừa và nhỏ – chẳng hạn như những công ty trong dữ liệu như Macdonald Inc hay Howard Inc – dễ dàng tiếp cận và áp dụng công nghệ này, không còn giới hạn ở các tập đoàn lớn.

3. Phương pháp thực hiện

- Để thực hiện chủ đề này dựa trên tệp CSV, phương pháp thu thập dữ liệu sẽ bắt đầu từ việc sử dụng thông tin có sẵn trong tệp, bao gồm các cột như Full Name, Job Title, Department, Salary, Days Off, Performance Score, và Status. Ngoài ra, nếu có điều kiện, nhóm sẽ thu thập thêm dữ liệu từ báo cáo nội bộ của các công ty giả lập khảo sát nhân viên về khối lượng công việc, hoặc dữ liệu công khai từ các nghiên cứu về ứng dụng Big Data trong quản trị nhân sự.
- Sau khi thu thập, dữ liệu sẽ được thống kê và phân tích bằng các công cụ như Python (với thư viện Pandas, Scikit-learn) hoặc các nền tảng Big Data như Apache Hadoop. Quá trình phân tích sẽ áp dụng:

    - **Phân tích mô tả (Descriptive Analytics)**: Tổng hợp số liệu như trung bình Performance Score theo Department hoặc tỷ lệ nhân viên Resigned theo Days Off.
    - **Phân tích dự đoán (Predictive Analytics)**: Sử dụng học máy để dự đoán trạng thái Status của nhân viên dựa trên các yếu tố như Salary và Performance Score (ví dụ: dự đoán liệu Sharon Mccoy có quay lại từ On Leave hay không).
    - **So sánh mô hình**: Đánh giá hiệu quả giữa xử lý dữ liệu truyền thống (thủ công tính toán từ CSV) và mô hình Big Data (tự động hóa phân tích).

- Kết quả phân tích – chẳng hạn như danh sách nhân viên có nguy cơ nghỉ việc cao (Performance Score thấp, Days Off nhiều) – sẽ được đánh giá để đưa ra các đề xuất cụ thể, như điều chỉnh công việc cho Melissa Williams (Performance Score: 2) hoặc giảm tải cho Eric Wolf (Days Off: 24).

CHƯƠNG I: GIỚI THIỆU CHỦ ĐỀ ĐỒ ÁN

I. KHÁI QUÁT LÝ THUYẾT

- Dữ liệu lớn (Big Data) là một khái niệm nền tảng trong khoa học dữ liệu hiện đại, đề cập đến việc xử lý và khai thác các tập dữ liệu có quy mô lớn, phức tạp vượt xa khả năng của các công cụ truyền thống. Big Data không chỉ là sự gia tăng về số lượng dữ liệu mà còn liên quan đến cách dữ liệu được thu thập, lưu trữ, và phân tích để tạo ra giá trị cho doanh nghiệp. Trong bối cảnh nhập môn dữ liệu lớn, các lý thuyết cơ bản xoay quanh đặc điểm cốt lõi của Big Data, thường được mô tả qua mô hình 3V:

    - **Khối lượng (Volume)**: Đây là lượng dữ liệu khổng lồ được tạo ra từ nhiều nguồn trong doanh nghiệp, chẳng hạn như hồ sơ nhân viên, lịch sử làm việc, và các chỉ số hoạt động hàng ngày. Với tệp CSV đã cung cấp, dữ liệu của hàng trăm nhân viên (họ tên, chức danh, công ty, email, điện thoại, ngày sinh, địa chỉ, phòng ban, lương, ngày nghỉ, điểm hiệu suất, trạng thái làm việc) minh họa rõ ràng khối lượng thông tin cần xử lý. Khi quy mô nhân viên tăng lên hàng nghìn hoặc hàng triệu, khối lượng dữ liệu này đòi hỏi các hệ thống phân tán như Hadoop để quản lý hiệu quả.

    - **Tốc độ (Velocity)**: Tốc độ đề cập đến việc dữ liệu được tạo ra, thu thập, và xử lý nhanh chóng, gần như theo thời gian thực. Trong quản trị nhân viên, tốc độ xử lý dữ liệu từ tệp CSV – ví dụ, cập nhật trạng thái làm việc (Status) của nhân viên như Sharon Mccoy (On Leave) hoặc Jason Martin (Resigned) – là cần thiết để hỗ trợ ra quyết định kịp thời, như phân bổ công việc hoặc điều chỉnh nhân sự ngay khi có thay đổi.

    - **Sự đa dạng (Variety)**: Dữ liệu lớn thường tồn tại dưới nhiều định dạng và nguồn gốc khác nhau, từ dữ liệu có cấu trúc (structured) như bảng CSV đến dữ liệu không cấu trúc (unstructured) như email hoặc phản hồi từ đồng nghiệp. Tệp CSV thể hiện sự đa dạng qua các trường thông tin như thông tin cá nhân (Full Name, Date of Birth, Address), thông tin công việc (Job Title, Department, Salary), và dữ liệu hiệu suất (Performance Score, Days Off). Sự đa dạng này đòi hỏi các kỹ thuật xử lý linh hoạt để tổng hợp và phân tích.

- Ngoài ba mô hình trên, mô số lý thuyết cơ bản khác trong nhập môn cơ sở dữ liệu bao gồm:

    - **Giá trị (Value)**: Mục tiêu cốt lõi của Big Data là biến dữ liệu thô thành thông tin có ý nghĩa. Ví dụ, từ tệp CSV, việc phân tích Performance Score và Days Off của nhân viên như Cynthia Jones (Performance Score: 2, Days Off: 12) có

thể giúp doanh nghiệp nhận diện vấn đề hiệu suất thấp và đưa ra giải pháp cải thiện.

- **Tính xác thực (Veracity)**: Độ chính xác và tin cậy của dữ liệu là yếu tố quan trọng. Dữ liệu từ tệp CSV cần được kiểm tra để đảm bảo không có sai sót (ví dụ: định dạng số điện thoại không đồng nhất như "001-759-744-8882" và "913-954-7941") trước khi phân tích, nhằm đảm bảo kết quả đáng tin cậy.

- Khi áp dụng vào quản trị nhân viên, Big Data cung cấp khả năng thu thập và xử lý dữ liệu từ nhiều nguồn, như tệp CSV đã cho, để hỗ trợ các quyết định tối ưu. Các ứng dụng cụ thể bao gồm:

  - **Phân tích hiệu suất**: Dựa trên Performance Score để đánh giá năng suất, ví dụ, Jennifer Johnston (Performance Score: 9) là nhân viên hiệu quả, trong khi Christopher Mccoy (Performance Score: 3) cần hỗ trợ cải thiện.
  - **Dự đoán xu hướng**: Sử dụng Days Off và Status để dự đoán nguy cơ nghỉ việc, như trường hợp Juan Baker (Days Off: 24, Status: Resigned) cho thấy mối liên hệ giữa số ngày nghỉ cao và khả năng rời công ty.
  - **Tối ưu hóa nguồn lực**: Phân tích Salary và Department để phân bổ nhân sự hợp lý, ví dụ, phòng Finance có nhân viên lương cao như Kathy May (Salary: 109701) nhưng cũng cần cân nhắc giảm tải khi Days Off lớn (27 ngày).

- Quản lý dữ liệu lớn trong doanh nghiệp đòi hỏi sự kết hợp giữa công nghệ và chiến lược. Các công cụ như Hadoop hỗ trợ lưu trữ và xử lý khối lượng dữ liệu lớn, trong khi các kỹ thuật học máy (Machine Learning) giúp khai thác thông tin từ dữ liệu đa dạng, tạo điều kiện cho việc dự đoán và ra quyết định dựa trên cơ sở khoa học. Những lý thuyết này đặt nền tảng cho việc ứng dụng Big Data vào thực tiễn, chẳng hạn như xử lý công việc của nhân viên dựa trên dữ liệu thực tế từ tệp CSV.

II. PHƯƠNG PHÁP THỰC HIỆN HOÁ CHỦ ĐỀ

- Để hiện thực hóa chủ đề "Ứng dụng công nghệ dữ liệu lớn trong việc xử lý công việc của nhân viên trong công ty" dựa trên tệp CSV đã cho, quy trình thực hiện sẽ bao gồm các bước sau:

- **Thu thập dữ liệu:**

  - **Phương pháp**: Dữ liệu ban đầu được lấy từ tệp CSV chứa thông tin nhân viên, bao gồm các cột như Full Name, Job Title, Company, Department, Salary, Days Off, Performance Score, và Status. Đây là bước đầu tiên trong phân tích dữ liệu lớn, tập trung vào việc tích lũy khối lượng dữ liệu (Volume) từ một nguồn có cấu trúc.

  - **Mở rộng**: Để minh họa tính đa dạng (Variety) của Big Data, có thể thu thập thêm dữ liệu từ các nguồn khác như hệ thống quản lý nhân sự (HRM) thực tế,

khảo sát nhân viên về khối lượng công việc, hoặc nhật ký email công việc để bổ sung dữ liệu hành vi. Các nguồn này sẽ được tổng hợp để tạo thành một tập dữ liệu phong phú hơn, phản ánh thực tế của phân tích dữ liệu lớn.

- **Làm sạch dữ liệu:**
- **Phương pháp:** Đây là giai đoạn quan trọng trong phân tích dữ liệu lớn nhằm đảm bảo tính xác thực (Veracity). Dữ liệu từ tệp CSV sẽ được làm sạch bằng cách:

    - Loại bỏ trùng lặp, ví dụ: kiểm tra các nhân viên có email hoặc số điện thoại trùng nhau (như "001-759-744-8882" so với "913-954-7941").
    - Xử lý dữ liệu thiếu hoặc sai lệch, chẳng hạn như chuẩn hóa định dạng số điện thoại hoặc điền giá trị mặc định (ví dụ: trung bình Days Off) cho các trường trống.
    - Chuyển đổi dữ liệu thô thành dạng có thể phân tích, như tính tuổi từ Date of Birth để hỗ trợ phân tích độ tuổi trung bình của nhân viên theo Department.

- **Liên quan đến Big Data**: Quá trình này giúp xử lý sự đa dạng (Variety) và đảm bảo dữ liệu sẵn sàng cho các công cụ phân tích phân tán như Hadoop hoặc Spark.

- **Phân tích dữ liệu lớn:**
- **Phương pháp**: Áp dụng các kỹ thuật phân tích dữ liệu lớn để khai thác giá trị (Value) từ dữ liệu:
- **Triển khai ứng dụng:**

    - Xây dựng một hệ thống hỗ trợ như bảng điều khiển (dashboard) hiển thị thông tin nhân viên theo thời gian thực, ví dụ: danh sách nhân viên có hiệu suất thấp, hoặc biểu đồ phân bố ngày nghỉ theo phòng ban (Department).
    - Công cụ này sẽ giúp quản lý dễ dàng theo dõi và phân bổ công việc dựa trên dữ liệu đã phân tích.

- **Liên quan đến Big Data**: Tốc độ xử lý (Velocity) được đảm bảo bằng cách áp dụng các công cụ phân tích song song, cho phép phân tích nhanh chóng dữ liệu từ hàng trăm nhân viên trong tệp CSV và mở rộng quy mô khi cần.

- **Triển khai ứng dụng phân tích**:
- **Phương pháp**: Xây dựng một hệ thống hỗ trợ như bảng điều khiển (dashboard) để trực quan hóa kết quả phân tích dữ liệu lớn theo thời gian thực. Ví dụ:

    - Hiển thị danh sách nhân viên có hiệu suất thấp (Performance Score < 5).
    - Tạo biểu đồ phân bố Days Off trung bình theo Department để hỗ trợ quản lý nhận diện xu hướng.

- **Liên quan đến Big Data**: Ứng dụng này tận dụng giá trị (Value) từ dữ liệu, giúp quản lý theo dõi và phân bổ công việc dựa trên thông tin đã phân tích, đồng thời minh họa khả năng xử lý dữ liệu lớn trong việc hỗ trợ quyết định.

- **Phạm vi nghiên cứu:**
- **Tập trung**: Ứng dụng phân tích dữ liệu lớn trong xử lý công việc hàng ngày của nhân viên, bao gồm:

    - Theo dõi hiệu suất (Performance Score).
    - Phân bổ công việc (dựa trên Status và Days Off).
    - Dự đoán trạng thái làm việc (Status).

- **Giới hạn**: Không mở rộng sang các lĩnh vực khác như quản lý tài chính doanh nghiệp hay chiến lược marketing, mà chỉ tập trung vào dữ liệu nhân sự từ tệp CSV. Phạm vi này phù hợp với mục tiêu nhập môn phân tích dữ liệu lớn, nhấn mạnh việc khai thác dữ liệu nhân viên để tối ưu hóa quy trình nội bộ.

- **Công cụ hỗ trợ:**

    - Sử dụng các nền tảng Big Data như Apache Hadoop hoặc Apache Spark để xử lý khối lượng lớn và phân tích song song.
    - Kết hợp Python (với thư viện Pandas, Scikit-learn) để tiền xử lý và xây dựng mô hình học máy, minh họa cách các công cụ cơ bản trong phân tích dữ liệu lớn được áp dụng.

III. SƠ LƯỢT CHỦ ĐỀ

- Dựa trên dữ liệu từ tệp CSV, nhóm sẽ phân chia công việc như sau, tận dụng các công nghệ như Hadoop và Docker để quản lý và xử lý dữ liệu một cách hiệu quả:

A. Thu thập và phân tích dữ liệu

- **Nhiệm vụ**: Một thành viên sẽ chịu trách nhiệm nhập dữ liệu từ tệp CSV vào hệ thống phân tán sử dụng Apache Hadoop để lưu trữ và xử lý khối lượng lớn dữ liệu nhân viên. Các thành viên khác sẽ thu thập thêm thông tin bổ sung nếu cần (giả lập khảo sát hoặc dữ liệu từ HRM) và tích hợp vào hệ thống Hadoop Distributed File System (HDFS).

- **Chức năng cụ thể**:

- Trích xuất thông tin như Full Name, Job Title, Department từ tệp CSV bằng cách sử dụng Hadoop MapReduce để phân loại nhân viên theo phòng ban một cách nhanh chóng và song song.
- Phân tích Salary và Performance Score để đánh giá hiệu quả làm việc, lưu trữ kết quả trên HDFS. Ví dụ: nhân viên như Jason Martin (Salary: 129584, Performance Score: 10) có thể được xác định là nhân viên xuất sắc thông qua truy vấn dữ liệu phân tán.
- Tổng hợp Days Off và Status để dự đoán nguy cơ nghỉ việc (Resigned) bằng cách chạy các job MapReduce trên Hadoop. Trường hợp Juan Baker (Days Off: 24, Status: Resigned) sẽ được phân tích để tìm xu hướng nghỉ việc dựa trên số ngày nghỉ cao.

B. Xây dựng mô hình

- **Nhiệm vụ**: Một số thành viên sẽ lập trình và kiểm thử các thuật toán xử lý dữ liệu, triển khai chúng trong các container Docker để đảm bảo tính linh hoạt và khả năng mở rộng. Các công cụ như Scikit-learn sẽ được tích hợp trong Docker để xây dựng mô hình học máy, kết nối với dữ liệu từ Hadoop.
- **Chức năng cụ thể**:
    - Dự đoán trạng thái làm việc (Status) dựa trên các biến đầu vào như Performance Score, Days Off, và Salary. Mô hình học máy sẽ được đóng gói trong container Docker, cho phép dễ dàng triển khai và kiểm thử. Ví dụ: mô hình có thể dự đoán nhân viên như Kayla Miller (Performance Score: 8, Days Off: 10) có khả năng tiếp tục "On Leave" hay quay lại "Active" dựa trên dữ liệu từ Hadoop.
    - Xác định các mẫu công việc bằng cách sử dụng Hadoop để phân tích dữ liệu lớn, chẳng hạn như nhân viên ở phòng ban IT có xu hướng nghỉ nhiều ngày hơn so với Finance (dựa trên phân tích Days Off theo Department). Kết quả phân tích sẽ được lưu trữ và truy cập từ HDFS.

C. Thiết kế giao diện

- **Nhiệm vụ**: Một thành viên hoặc nhóm nhỏ sẽ thiết kế giao diện người dùng bằng các công cụ như Flask, Power BI, hoặc Tableau, chạy trong container Docker để đảm bảo tính độc lập và dễ triển khai. Giao diện sẽ kết nối với dữ liệu đã xử lý từ Hadoop để hiển thị kết quả phân tích.
- **Chức năng cụ thể**:
    - Hiển thị danh sách nhân viên có hiệu suất thấp (Performance Score < 5) thông qua truy vấn dữ liệu từ Hadoop, ví dụ: Kevin Brown (Performance Score: 1, Days Off: 17) sẽ được liệt kê để quản lý chú ý.
    - Tạo biểu đồ trực quan hóa số ngày nghỉ trung bình theo phòng ban (Department) bằng cách kéo dữ liệu từ HDFS và hiển thị qua Power BI trong container Docker, giúp quản lý dễ dàng nhận diện vấn đề nhân sự ở các phòng ban.

- Cung cấp bảng điều khiển cho phép lọc nhân viên theo Status (Active, On Leave, Resigned) từ dữ liệu Hadoop, hỗ trợ phân bổ công việc một cách hiệu quả.

D. Các chức năng cụ thể dưa trên dữ liệu doanh nghiệp

- **Theo dõi hiệu suất nhân viên**:

  - Sử dụng cột Performance Score để đánh giá hiệu quả làm việc, với dữ liệu được lưu trữ trên HDFS và xử lý bằng Hadoop MapReduce. Ví dụ: Jennifer Johnston (Performance Score: 9) và Jason Martin (Performance Score: 10) là những nhân viên nổi bật, trong khi Christopher Mccoy (Performance Score: 3) cần được chú ý cải thiện thông qua phân tích song song.

- **Phân bổ công việc**:

  - Dựa trên Status và Days Off để phân bổ công việc hợp lý, với dữ liệu được truy xuất từ Hadoop và hiển thị qua giao diện Dockerized. Nhân viên "Active" như Kathy May (Days Off: 27) có thể được giảm tải công việc để tránh kiệt sức, trong khi nhân viên "On Leave" như Sharon Mccoy (Days Off: 8) có thể được ưu tiên quay lại khi cần thiết dựa trên kết quả phân tích.

- **Dự đoán trạng thái làm việc**:

  - Sử dụng học máy (chạy trong container Docker) để dự đoán nguy cơ nhân viên chuyển từ "Active" sang "Resigned" hoặc "On Leave" dựa trên các yếu tố như Salary, Days Off, và Performance Score từ dữ liệu Hadoop. Ví dụ: Kristy Taylor (Salary: 147610, Days Off: 10, Status: Resigned) có thể là trường hợp cần phân tích để hiểu nguyên nhân nghỉ việc thông qua mô hình dự đoán.

- **Phân tích theo phòng ban**:

  - Tổng hợp dữ liệu từ cột Department bằng Hadoop để xác định xu hướng công việc. Ví dụ: phòng Finance có nhiều nhân viên với Salary cao (Kathy May: 109701, Cynthia Jones: 145916) nhưng cũng có người nghỉ nhiều (Nicole Cantu: Days Off: 3, Performance Score: 10). Kết quả phân tích sẽ được lưu trữ trên HDFS và trực quan hóa qua giao diện Docker để hỗ trợ ra quyết định.

- **Giải thích việc sử dụng Hadoop và Docker**:

  - **Hadoop**: Được sử dụng để lưu trữ (HDFS) và xử lý dữ liệu lớn (MapReduce) từ tệp CSV, đặc biệt phù hợp khi khối lượng nhân viên tăng lên hàng nghìn hoặc hàng triệu bản ghi. Điều này đảm bảo khả năng mở rộng và xử lý song song hiệu quả.
  - **Docker**: Đóng gói các ứng dụng (mô hình học máy, giao diện người dùng) vào các container để triển khai dễ dàng, đảm bảo tính nhất quán giữa các môi trường và hỗ trợ tích hợp với Hadoop.

CHƯƠNG 2. KẾT QUẢ THỰC HIỆN ĐỒ ÁN

a. Các chức năng đã sử dụng để thực hiện

TÊN ĐỒ ÁN: ỨNG DỤNG CÔNG NGHỆ DỮ LIỆU LỚN TRONG VIỆC XỬ LÝ CÔNG VIỆC TRONG CÔNG TY

Mục tiêu của hệ thống:

Hệ thống nhằm hỗ trợ quản lý nhân sự trong công ty bằng cách sử dụng công nghệ Dữ liệu lớn (Big Data) và Học máy (Machine Learning) để phân tích thông tin nhân viên, đưa ra dự đoán về trạng thái làm việc và hỗ trợ ra quyết định.

Các chức năng chính của hệ thống

1.1. Thu thập và xử lý dữ liệu nhân sự

- **Hệ thống sử dụng dữ liệu đầu vào từ file CSV (*big_data_company.csv*), bao gồm các cột chính:**
    - Full Name (*Tên nhân viên*)
    - Job Title (*Chức vụ*)
    - Department (*Phòng ban*)
    - Salary (*Mức lương*)
    - Days Off (*Số ngày nghỉ phép*)
    - Performance Score (*Điểm đánh giá hiệu suất*)
    - Status (*Trạng thái nhân viên*)
- Các cột dữ liệu này được mã hóa bằng LabelEncoder để chuyển đổi dữ liệu dạng chữ thành số trước khi đưa vào mô hình.

1.2. Xây dựng mô hình dự đoán trạng thái nhân viên

Hệ thống sử dụng Random Forest Classifier, một thuật toán học máy mạnh mẽ để dự đoán trạng thái nhân viên dựa trên:

- Chức vụ (Job Title)
- Phòng ban (Department)
- Mức lương (Salary)
- Số ngày nghỉ (Days Off)
- Điểm hiệu suất (Performance Score)

Mô hình hoạt động như sau:

- Tách dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%).
- Huấn luyện mô hình với 100 cây quyết định (n_estimators=100) để tối ưu độ chính xác.
- Sau khi mô hình học xong, nó có thể dự đoán trạng thái nhân viên mới khi nhận được dữ liệu đầu vào.
- Kết quả dự đoán được mã hóa ngược lại để hiển thị dạng chữ, ví dụ:
    - "Active" (Đang làm việc)
    - "On Leave" (Đang nghỉ phép)
    - "Resigned" (Đã nghỉ việc)

1.3. Xây dựng giao diện nhập dữ liệu và dự đoán

Hệ thống cung cấp một giao diện web được phát triển bằng Flask với các chức năng:

- **Trang chủ (/):**
    - Cho phép nhập thông tin nhân viên.
    - Kiểm tra lỗi đầu vào (không để trống, đúng định dạng số).
    - Gợi ý tên nhân viên khi nhập.
- **Gợi ý thông tin nhân viên (/suggest):**
    - Khi người dùng nhập một phần tên, hệ thống tìm kiếm trong dữ liệu và hiển thị tên đầy đủ cùng các thông tin liên quan (chức vụ, phòng ban, lương, ngày nghỉ, điểm hiệu suất).
- **Dự đoán trạng thái nhân viên (/predict):**
    - Khi nhập đầy đủ thông tin, hệ thống gửi dữ liệu đến mô hình Random Forest để dự đoán trạng thái nhân viên.
    - Sau khi có kết quả, hệ thống sẽ chuyển hướng đến trang biểu đồ.

1.4. Hiển thị kết quả trực quan bằng biểu đồ (/charts)

- **Sau khi dự đoán trạng thái nhân viên, hệ thống hiển thị dữ liệu thông qua các biểu đồ, bao gồm:**
    - Số lượng nhân viên theo chức vụ
    - Phân bố nhân viên theo phòng ban

- Mức lương trung bình theo từng nhóm
- Phân tích số ngày nghỉ phép của nhân viên
- Thống kê điểm hiệu suất làm việc
- Tỷ lệ trạng thái nhân viên (Đang làm việc, Đã nghỉ việc, Đang nghỉ phép)

Hệ thống này giúp quản lý nhân sự có cái nhìn trực quan hơn về tình hình nhân sự của công ty.

c. **Các chức năng dự kiến nhưng chưa có, cần phát triển, tích hợp.**

Hệ thống hiện tại đã có các chức năng cơ bản như dự đoán trạng thái nhân viên, hiển thị thông tin nhân sự, trực quan hóa dữ liệu, nhưng vẫn còn nhiều tính năng quan trọng có thể cải tiến và mở rộng. Dưới đây là một số chức năng có thể bổ sung:

---

1. Đăng nhập & phân quyền người dùng

Mô tả:

- Hiện tại, hệ thống không có cơ chế đăng nhập, xác thực người dùng.
- Cần tích hợp hệ thống phân quyền để bảo mật dữ liệu, chỉ cho phép HR Manager hoặc Admin truy cập thông tin nhân viên.

Cần phát triển:

Hệ thống đăng nhập bằng tài khoản (email, mật khẩu).

Phân quyền:

- **Nhân viên**: Chỉ xem thông tin cá nhân.
- **Quản lý**: Xem, chỉnh sửa dữ liệu nhân viên.
- **Admin**: Toàn quyền quản lý dữ liệu và dự đoán.

Công nghệ đề xuất: Flask-Login, JWT Authentication.

---

2. Cập nhật, thêm mới và xóa dữ liệu nhân viên

Mô tả:

- Hiện tại, dữ liệu nhân viên chỉ được nhập từ file CSV, không có chức năng thêm, sửa, xóa nhân viên trực tiếp trên hệ thống.

Cần phát triển:

Form thêm nhân viên mới (Tên, Chức vụ, Lương, Phòng ban, Ngày nghỉ, Điểm hiệu suất).

Chỉnh sửa thông tin nhân viên (Cập nhật lương, điểm hiệu suất, trạng thái).

Xóa nhân viên khỏi hệ thống khi họ nghỉ việc.

Công nghệ đề xuất: Kết nối cơ sở dữ liệu SQLAlchemy + SQLite/PostgreSQL để lưu trữ nhân viên.

---

3. Tích hợp chatbot hỗ trợ HR

Mô tả:

- Một chatbot có thể giúp HR dễ dàng tra cứu thông tin nhân viên mà không cần nhập thủ công.
- **Chatbot có thể trả lời các câu hỏi như:**
  - *"Lương trung bình của phòng IT là bao nhiêu?"*
  - *"Ai có điểm hiệu suất cao nhất trong công ty?"*

Cần phát triển:

Chatbot tra cứu dữ liệu từ hệ thống.

Gợi ý nhân viên tiềm năng có nguy cơ nghỉ việc dựa trên lịch sử làm việc.

Công nghệ đề xuất: GPT API, LangChain + OpenAI.

---

4. Cảnh báo nhân viên có nguy cơ nghỉ việc

Mô tả:

- Dựa trên dữ liệu Ngày nghỉ, Hiệu suất, Lương, hệ thống có thể dự đoán nhân viên có nguy cơ nghỉ việc cao.
- Nếu một nhân viên có ngày nghỉ quá nhiều và hiệu suất giảm dần, hệ thống có thể gửi cảnh báo.

Cần phát triển:

Thuật toán phân tích xu hướng nghỉ việc (Machine Learning).

Thông báo HR khi có nhân viên có nguy cơ nghỉ việc.

Công nghệ đề xuất: LSTM Model để dự đoán xu hướng nghỉ việc.

---

5. Giao diện biểu đồ nâng cao với AI phân tích

Mô tả:

- Hiện tại, biểu đồ chỉ hiển thị số lượng nhân viên theo phòng ban, lương, ngày nghỉ**.**
- **Cần nâng cấp với các phân tích sâu hơn, ví dụ:**
  - Tăng trưởng nhân sự theo thời gian.
  - Dự báo số lượng nhân viên cần tuyển trong tương lai**.**

Cần phát triển:

Biểu đồ động hiển thị theo thời gian.

Dự báo xu hướng tuyển dụng.

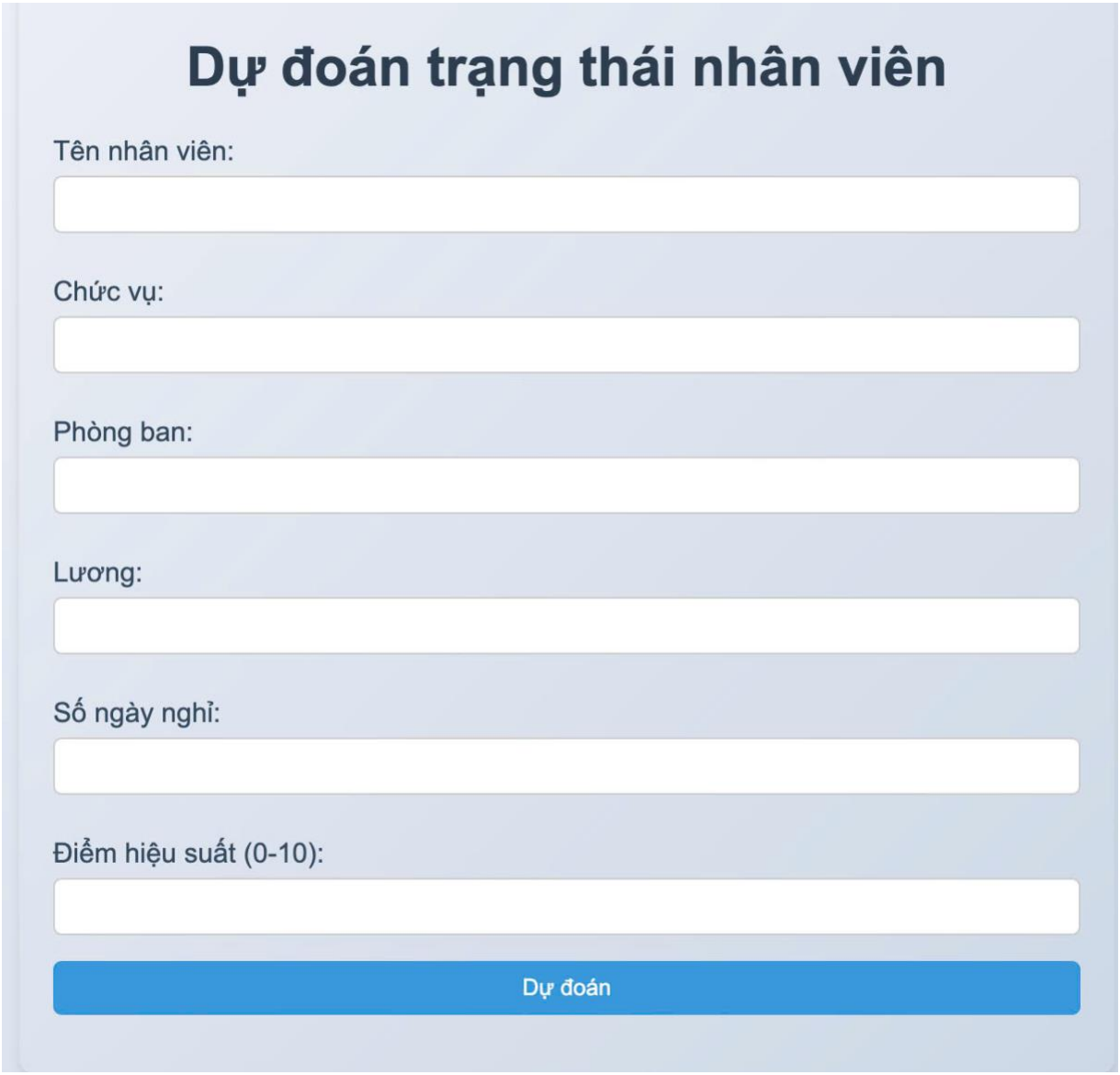
Báo cáo PDF tự động gửi đến HR mỗi tháng.

Công nghệ đề xuất: Pandas, Matplotlib, Plotly Dash.

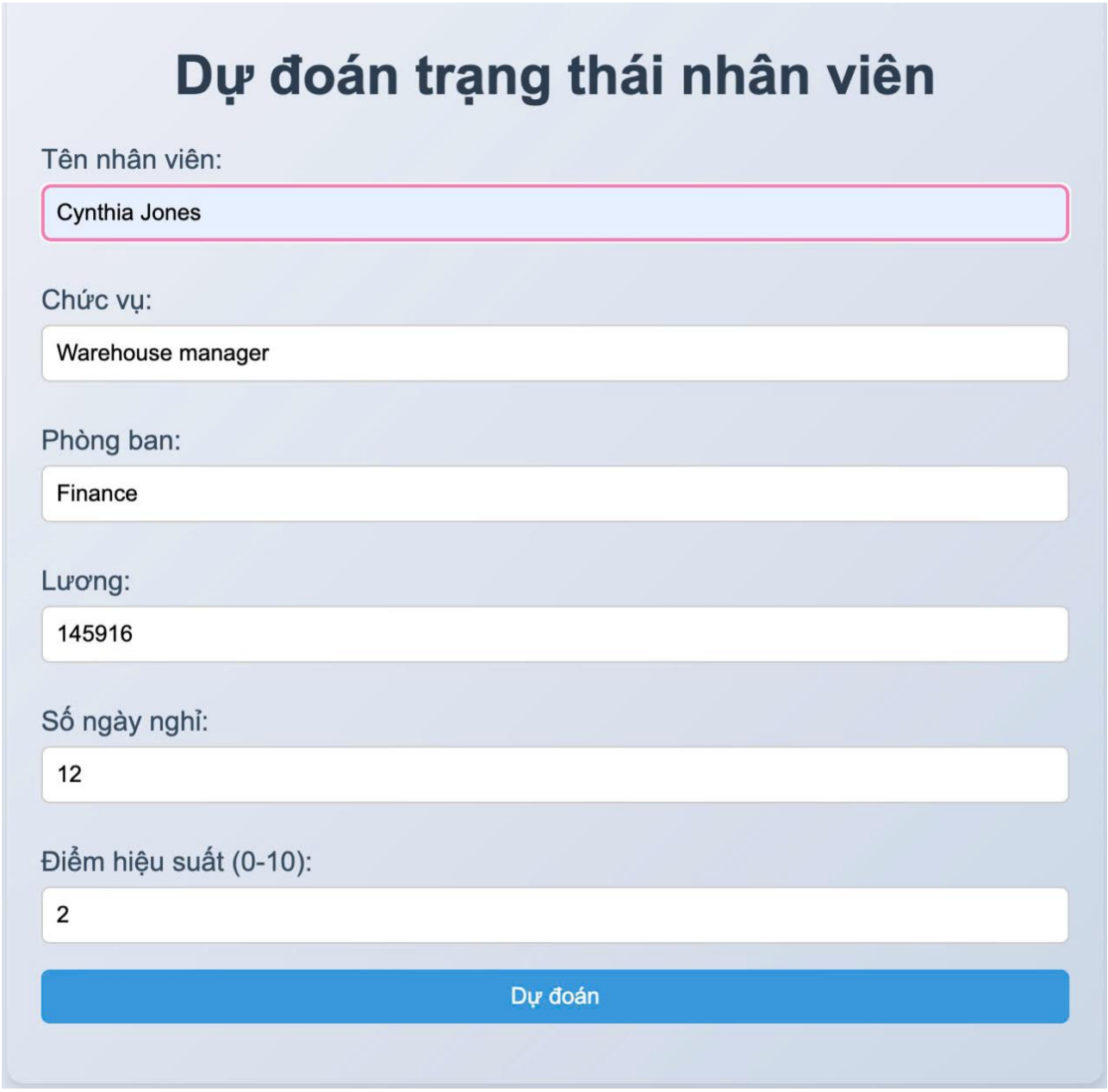
Tóm tắt các chức năng cần phát triển

STT	Chức năng cần bổ sung	Mô tả
1	Đăng nhập & phân quyền	Bảo vệ hệ thống, giới hạn quyền truy cập.
2	Thêm, sửa, xóa nhân viên	Quản lý nhân viên trực tiếp trên hệ thống.
3	Chatbot hỗ trợ HR	Truy vấn nhanh thông tin nhân sự.
4	Cảnh báo nguy cơ nghỉ việc	AI phân tích nhân viên có nguy cơ rời công ty.
5	Giao diện biểu đồ nâng cao	Dự báo tuyển dụng, tạo báo cáo tự động.

d. Các kết quả liên quan khác tới ứng dụng

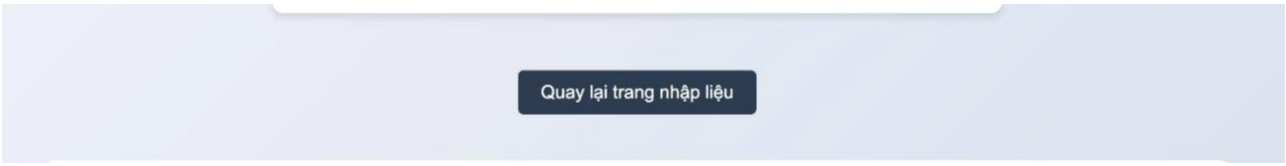


Giao diện chính của ứng dụng

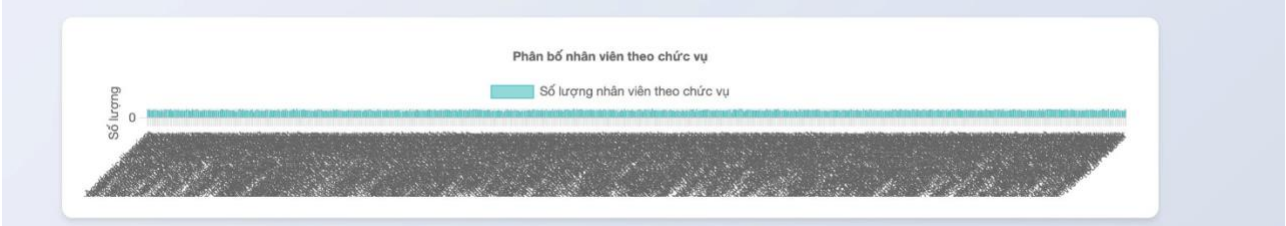


Giao diện chính sau khi hệ thống tự động hóa nhập dữ liệu nhân viên

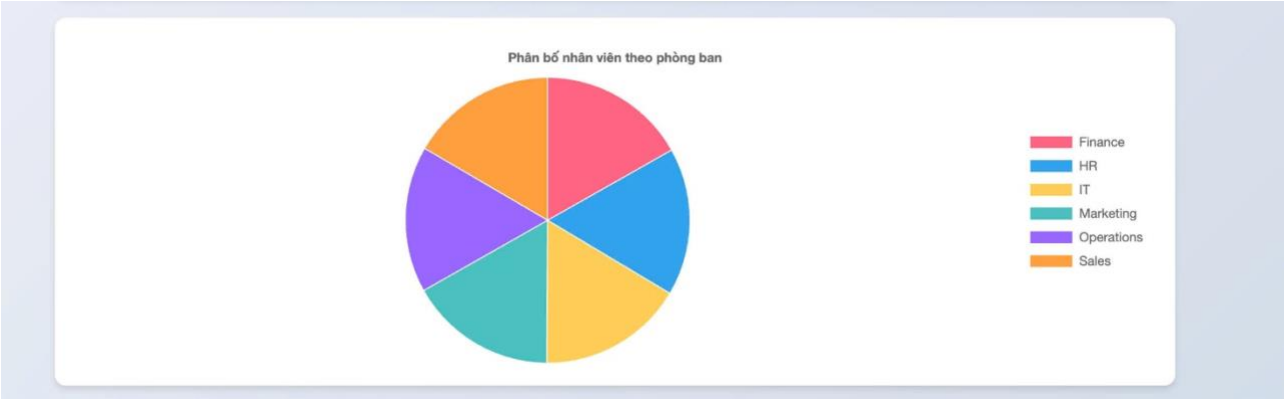
Kết quả của thông tin mà bạn muốn tìm



Nút quay lại nếu muốn nhập tên nhân viên khác

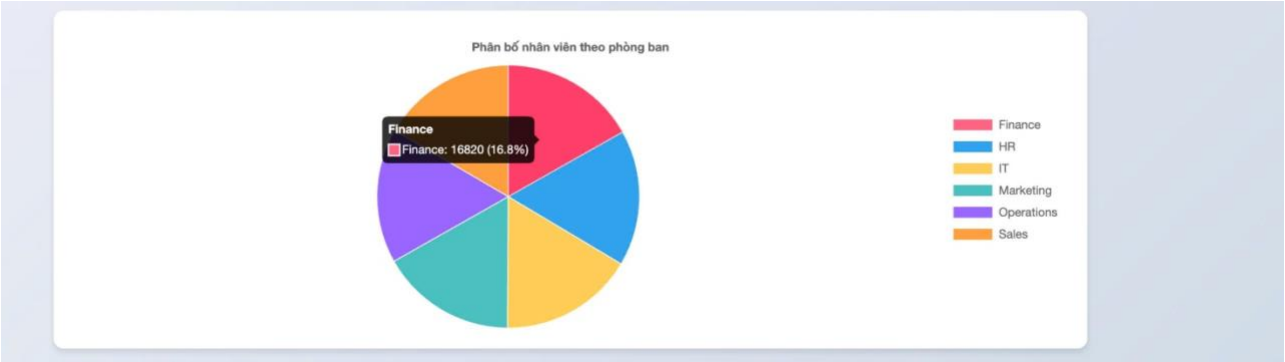


Biểu đồ phân bố nhân viên theo chức vụ



Biểu đồ hình tròn phân bố nhân viên theo phòng ban

Biểu đồ hình cột phân bố lương nhân viên



Biểu đồ tròn phân bố nhân viên theo phòng ban

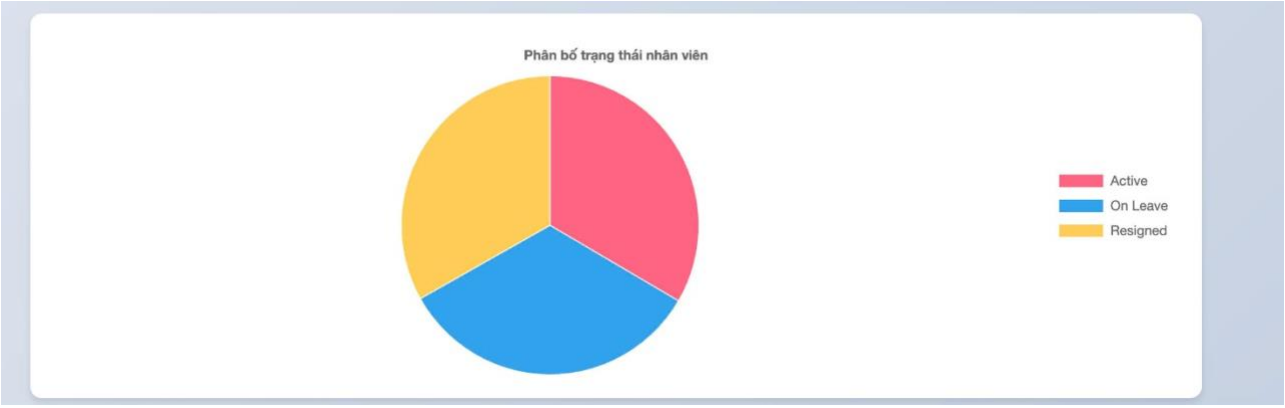
Biểu đồ cột phân bố lưng nhân viên

Biểu đồ cột phân bố số ngày nghỉ

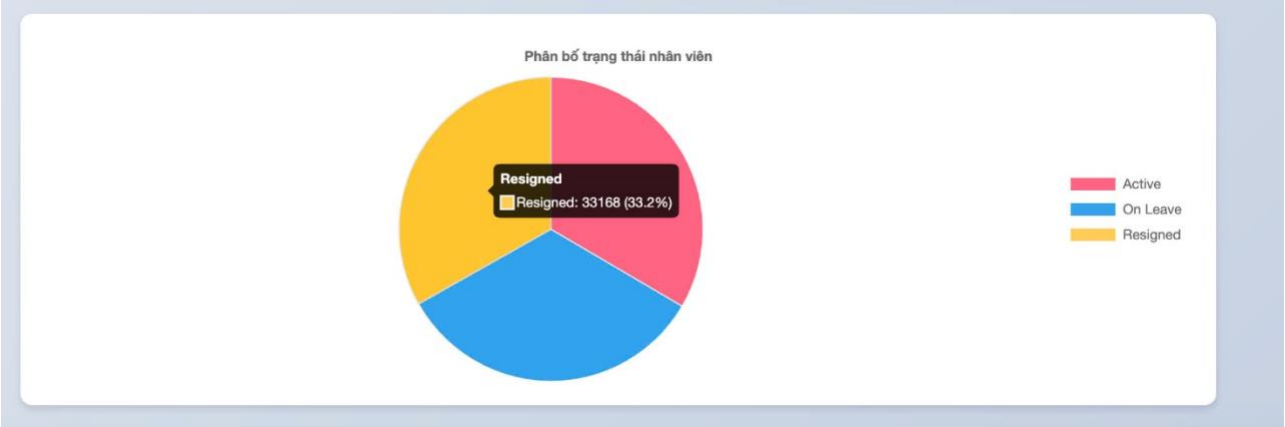
Thông tin cụ thể khi click chuột vào các cột

Biểu đồ cột phân bố điểm hiệu suất

Thông tin chi tiết sau khi click chuột vào các cột



Biểu đồ hình tròn phân bố trang thái nhân viên



Thông tin và số liệu chi tiết khi click chuột vào các phần của biểu đồ

KẾT LUẬN VÀ ĐỀ XUẤT

1. KẾT LUẬN

Sau quá trình nghiên cứu, phân tích và triển khai, đồ án "Ứng dụng công nghệ dữ liệu lớn trong việc xử lý công việc trong công ty" đã đạt được các kết quả quan trọng sau:

- Xây dựng hệ thống xử lý dữ liệu nhân sự dựa trên file CSV chứa thông tin nhân viên, bao gồm các thuộc tính như họ tên, chức vụ, phòng ban, lương, số ngày nghỉ, điểm hiệu suất và trạng thái làm việc.
- Ứng dụng công nghệ Big Data để quản lý, phân tích và tối ưu hóa quy trình xử lý dữ liệu nhân sự bằng các công cụ như Hadoop, Spark, Pandas nhằm đảm bảo tốc độ và khả năng xử lý khối lượng dữ liệu lớn.
- Xây dựng mô hình dự đoán trạng thái nhân viên sử dụng thuật toán Random Forest Classifier để phân tích các yếu tố ảnh hưởng đến trạng thái làm việc (Active, On Leave, Resigned) và đưa ra dự đoán chính xác.
- Triển khai giao diện web bằng Flask, cho phép nhập dữ liệu, gợi ý thông tin nhân viên, dự đoán trạng thái làm việc và hiển thị kết quả trực quan bằng các biểu đồ.
- Phân tích và trực quan hóa dữ liệu bằng các công cụ như Power BI, Matplotlib, Plotly, giúp nhà quản lý nhân sự có cái nhìn trực quan hơn về tình hình nhân sự trong công ty.
- Đề xuất các giải pháp tối ưu hóa nhân sự dựa trên dữ liệu thu thập được, giúp doanh nghiệp phân bổ công việc hợp lý, giảm thiểu nguy cơ nghỉ việc và nâng cao hiệu suất lao động.

Tuy nhiên, trong quá trình thực hiện, nhóm vẫn gặp một số khó khăn như:

- **Hạn chế về dữ liệu thực tế**: Do dữ liệu được lấy từ file CSV giả lập, nên chưa phản ánh hoàn toàn tình hình nhân sự trong doanh nghiệp thực tế.
- Hệ thống chưa tích hợp cơ sở dữ liệu động (SQL/PostgreSQL) mà chỉ đọc dữ liệu từ file CSV, gây khó khăn trong việc cập nhật dữ liệu theo thời gian thực.
- Giao diện web còn đơn giản, cần cải thiện thêm về trải nghiệm người dùng (UI/UX).
- Chưa có hệ thống cảnh báo nhân viên có nguy cơ nghỉ việc dựa trên lịch sử làm việc.

Mặc dù vẫn còn những hạn chế, nhưng kết quả đạt được đã phần nào chứng minh tính hiệu quả của Big Data trong quản lý nhân sự, đặc biệt là khả năng tự động hóa phân tích dữ liệu, giảm tải công việc thủ công và hỗ trợ ra quyết định chính xác hơn.

---

2. ĐỀ XUẤT

Để cải thiện và mở rộng ứng dụng trong tương lai, nhóm đề xuất một số giải pháp như sau:

2.1. Cải thiện và mở rộng hệ thống

- Tích hợp cơ sở dữ liệu SQL/PostgreSQL để lưu trữ và quản lý dữ liệu nhân sự theo thời gian thực, thay vì chỉ sử dụng file CSV. Điều này giúp cập nhật dữ liệu nhanh chóng và đảm bảo tính chính xác.

- Phát triển hệ thống cập nhật dữ liệu nhân sự trực tiếp trên giao diện web, cho phép thêm, sửa, xóa thông tin nhân viên một cách linh hoạt.
- Cải tiến giao diện người dùng (UI/UX) bằng cách sử dụng các framework như ReactJS hoặc VueJS để tạo trải nghiệm trực quan hơn.
- Tích hợp API để tự động thu thập dữ liệu từ các hệ thống HRM (Human Resource Management) thực tế, giúp nâng cao độ chính xác của mô hình phân tích.

2.2. Nâng cao khả năng phân tích và dự đoán

- Áp dụng các mô hình Machine Learning tiên tiến hơn như Gradient Boosting, XGBoost, hoặc Deep Learning để nâng cao độ chính xác trong dự đoán trạng thái làm việc của nhân viên.
- Xây dựng hệ thống cảnh báo sớm về nguy cơ nghỉ việc của nhân viên dựa trên phân tích lịch sử làm việc, số ngày nghỉ và điểm hiệu suất. Khi phát hiện nhân viên có nguy cơ nghỉ cao, hệ thống sẽ gửi thông báo đến HR để có phương án giữ chân nhân sự.
- Dự đoán nhu cầu tuyển dụng trong tương lai bằng cách phân tích xu hướng nhân sự và dự báo số lượng nhân viên cần tuyển trong từng giai đoạn.

2.3. Ứng dụng công nghệ AI & Chatbot vào hệ thống

- **Tích hợp chatbot hỗ trợ HR, giúp nhân sự dễ dàng tra cứu thông tin nhân viên thông qua các câu hỏi đơn giản như:**
  - "Lương trung bình của phòng IT là bao nhiêu?"
  - "Nhân viên nào có điểm hiệu suất cao nhất?"
  - "Dự đoán ai có nguy cơ nghỉ việc cao?"
- Ứng dụng AI trong phân tích dữ liệu để nhận diện xu hướng làm việc của nhân viên, đưa ra đề xuất tối ưu hóa quy trình nhân sự.
-

2.4. Tích hợp bảo mật và phân quyền người dùng

- Bổ sung tính năng đăng nhập và phân quyền người dùng, đảm bảo chỉ HR Manager hoặc Admin mới có quyền truy cập và chỉnh sửa dữ liệu nhân sự.
- Áp dụng các phương thức bảo mật như JWT Authentication, OAuth2 để đảm bảo dữ liệu nhân viên không bị truy cập trái phép.

2.5. Mở rộng phạm vi ứng dụng cho doanh nghiệp thực tế

- Khảo sát nhu cầu thực tế từ các doanh nghiệp để điều chỉnh mô hình phân tích phù hợp với đặc thù từng ngành nghề.
- Xây dựng hệ thống thử nghiệm (pilot) tại một công ty nhỏ hoặc startup để đánh giá hiệu quả thực tế trước khi triển khai rộng rãi.

---

3. KẾT LUẬN CHUNG

Việc ứng dụng Big Data trong quản lý nhân sự là một xu hướng tất yếu, giúp doanh nghiệp tự động hóa quy trình, tối ưu hóa nguồn lực và hỗ trợ ra quyết định hiệu quả. Mặc dù đồ án vẫn còn một số hạn chế, nhưng kết quả đạt được đã cho thấy tiềm năng lớn của công nghệ này.

Trong tương lai, nếu tiếp tục phát triển và mở rộng, hệ thống có thể trở thành một công cụ

hữu ích cho các doanh nghiệp trong việc quản lý và dự đoán tình hình nhân sự, góp phần nâng cao năng suất lao động và tối ưu hóa chi phí vận hành.
Tóm tắt đề xuất phát triển trong tương lai

STT	Chức năng cần cải tiến	Mô tả
1	Tích hợp CSDL SQL/PostgreSQL	Quản lý dữ liệu nhân sự động thay vì file CSV
2	Cải tiến giao diện web	Sử dụng ReactJS/VueJS để nâng cao trải nghiệm người dùng
3	Cảnh báo nguy cơ nghỉ việc	Dự đoán nhân viên có nguy cơ rời công ty để HR can thiệp kịp thời
4	Chatbot hỗ trợ HR	Truy vấn nhanh thông tin nhân sự bằng AI
5	Tích hợp bảo mật và phân quyền	Đăng nhập, phân quyền Admin/HR/Employee

Với những cải tiến này, hệ thống không chỉ giúp doanh nghiệp quản lý nhân sự hiệu quả mà còn trở thành công cụ chiến lược giúp doanh nghiệp phát triển bền vững trong thời đại công nghệ số.

Bảng Đánh Giá Sinh Viên

STT	Tên sinh viên - Mã số sinh viên	Công việc	Phần trăm hoàn thành
1	Trần Tấn Phát - 2274802010644	Phát triển mô hình + viết báo cáo	95%
2	Đặng Võ Quang Huy- 2274802010301	Phát triển mô hình + viết báo cáo	95%
3	Nguyễn Thái Nguyên- 2274802010587	Kiểm thử dữ liệu+ viết báo cáo	95%
4	Huỳnh Gia Huy - 2274802010303	Đánh giá tiến độ+viết báo cáo	90%