

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
 TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

# DS317.P11 - KHÁM PHÁ DỮ LIỆU DOANH NGHIỆP

**Đề tài**

Dự đoán khả năng tốt nghiệp đúng hạn

Của sinh viên

Nhóm 9

Nguyễn Anh Phi 19522005

Nguyễn Phước Thắng 21522509

Trần Thành Hùng 20521374

Văn Ngọc Nhật Huy 20521418

Nguyễn Đinh Việt Thắng 20521898

TP. HỒ CHÍ MINH, tháng 10 năm 2024

**MỤC LỤC**

[**Danh mục hình ảnh - biểu đồ**](#_heading=h.gjdgxs) **4**

[**Chương 1. TỔNG QUAN ĐỀ TÀI**](#_heading=h.30j0zll) **5**

[1. Giới thiệu](#_heading=h.1fob9te) 5

[2. Phát biểu bài toán](#_heading=h.tyjcwt)5

3[. Đối tượng và phạm vi](#_heading=h.1t3h5sf) 5

4[. Mục tiêu](#_heading=h.4d34og8) 5

[**Chương 2. TIỀN XỬ LÝ DỮ LIỆU**](#_heading=h.17dp8vu) **6**

[1. Thông tin về dữ liệu](#_heading=h.lnxbz9) 6

[2. Tiền xử lý dữ liệu](#_heading=h.pkwqa1) 6

[3. Kết hợp dữ liệu](#_heading=h.39kk8xu) 6

[**Chương 3. PHÂN TÍCH DỮ LIỆU**](#_heading=h.1opuj5n) **6**

[1. Phân tích](#_heading=h.7y0jwsxawzz5) 6

[2. Thảo luận kết quả phân tích](#_heading=h.184mhaj) 6

[**Chương 4. XÂY DỰNG MÔ HÌNH ỨNG DỤNG**](#_heading=h.3s49zyc) **7**

[1. Chuẩn bị dữ liệu](#_heading=h.279ka65) 7

[2. Phương pháp đánh giá](#_heading=h.meukdy) 7

[3. Phương pháp thực nghiệm](#_heading=h.36ei31r) 7

3.1. SVM 7

[3.2. XGBoost](#_heading=h.45jfvxd) 7

[3.3. Random Forest](#_heading=h.lezc68eib901) 7

[3.4. Decision Tree](#_heading=h.huclb2394d0a) 7

[3.5. PCA](#_heading=h.nbesg9ovjlr5) 7

[3.6. Biến đổi dữ liệu](#_heading=h.vhb9pejahtvn) 7

[4. Kết quả thực nghiệm](#_heading=h.2koq656) 7

[4.1. Thực nghiệm cơ bản](#_heading=h.dmkjz2xqgnkq) 7

[4.2. Thực nghiệm mở rộng](#_heading=h.6o2ozh9cl4ij) 7

[**Chương 5. KẾT LUẬN**](#_heading=h.zu0gcz) **8**

[**TÀI LIỆU THAM KHẢO**](#_heading=h.1yyy98l) **9**

# 

# Danh mục hình ảnh - biểu đồ

# TỔNG QUAN ĐỀ TÀI

## Giới thiệu

* Ngày nay, lượng dữ liệu về sinh viên (UIT) ngày càng phong phú và chi tiết, được tích lũy qua nhiều năm và không ngừng gia tăng về số lượng. Dữ liệu này đóng vai trò quan trọng trong việc phản ánh hiệu suất học tập của sinh viên, giúp sinh viên điều chỉnh kế hoạch học tập một cách hiệu quả. Đồng thời cũng giúp nhà trường nắm bắt được tình trạng học tập của sinh viên để từ đó đưa ra các phương pháp hỗ trợ sinh viên trong học tập. Để giúp sinh viên và nhà trường dễ dàng đánh giá khả năng ra trường đúng hạn, các hệ thống dự đoán khả năng ra trường đúng hạn của sinh viên cũng được nghiên cứu và phát triển.
* Trong nghiên cứu này, nhóm em tập trung vào việc phát triển mô hình dự đoán khả năng ra trường đúng hạn của sinh viên dựa trên dữ liệu liên quan đến quá trình học tập và các yếu tố khác. Mục tiêu chính là phát triển mô hình học máy có thể dự đoán chính xác khả năng tốt nghiệp đúng hạn của sinh viên dựa trên các yếu tố như điểm trung bình từng kỳ, số tín chỉ đã hoàn thành, các cảnh cáo học vụ, bảo lưu, và các thông tin cá nhân khác. Xác định các yếu tố quan trọng nhất ảnh hưởng đến khả năng tốt nghiệp đúng hạn, giúp hiểu rõ hơn về những yếu tố nào có tác động lớn nhất đến thành công trong học tập của sinh viên. Phân loại và cảnh báo sớm cho các nhóm sinh viên có nguy cơ cao, từ đó đề xuất các biện pháp can thiệp và hỗ trợ phù hợp nhằm giảm thiểu rủi ro sinh viên bị chậm tốt nghiệp.

## Phát biểu bài toán

* Việc tốt nghiệp đúng hạn là một thước đo quan trọng trong đánh giá chất lượng đào tạo của một trường đại học. Tuy nhiên, có nhiều yếu tố ảnh hưởng đến tiến độ học tập của sinh viên, bao gồm cả yếu tố cá nhân và môi trường học tập. Với sự phát triển của công nghệ phân tích dữ liệu và học máy, bài toán dự đoán khả năng tốt nghiệp đúng hạn có thể được giải quyết dựa trên các thông tin về quá trình học tập của sinh viên.
* Bài toán đặt ra là xây dựng một hệ thống dự đoán nhằm nhận diện sớm những sinh viên có nguy cơ không thể hoàn thành chương trình học đúng hạn. Điều này giúp cố vấn học tập và nhà trường đưa ra các biện pháp hỗ trợ phù hợp, từ việc cá nhân hóa kế hoạch học tập, tổ chức các buổi tư vấn, cung cấp các khóa học bổ trợ, đến thiết kế chương trình học phù hợp hơn với năng lực của từng nhóm sinh viên.
* **Input** của bài toán nghiên cứu bao gồm:
* Điểm số: điểm TB toàn khóa, điểm TB tích lũy Tiến độ học tập: học kỳ đang học, năm học, số tín chỉ tích lũy, học kỳ còn lại
* Thông tin sinh viên: giới tính, lớp, khoa, hệ đào tạo, chuyên ngành 2
* **Output** của bài toán bao gồm: Có hoặc không tốt nghiệp đúng hạn

## Đối tượng và phạm vi

* **Đối tượng:** Bài toán dự đoán sinh viên có tốt nghiệp đúng hạn hay không dựa trên các dữ liệu về sinh viên như điểm trung bình chung, điểm trung bình học kỳ, chứng chỉ ngoại ngữ, cảnh báo học vụ. Bộ dữ liệu được thu thập từ thông tin sinh viên theo học tại Trường Đại học Công nghệ thông tin - ĐHQG TPHCM từ năm học 2012 đến 2016.

## Mục tiêu

* Bài toán hướng đến việc xây dựng một hệ thống dự đoán khả năng tốt nghiệp đúng hạn của sinh viên, đồng thời phân tích các yếu tố ảnh hưởng đến tiến độ học tập như điểm số, số tín chỉ, chứng chỉ bắt buộc, và các vấn đề cá nhân. Qua đó, các mô hình dự đoán sẽ được thiết kế và đánh giá để đảm bảo độ chính xác và tính ứng dụng cao.
* Kết quả dự đoán và phân tích sẽ cung cấp thông tin quan trọng cho nhà trường, giảng viên, và cố vấn học tập, giúp họ triển khai các biện pháp hỗ trợ hiệu quả nhằm nâng cao chất lượng học tập và tăng tỷ lệ tốt nghiệp đúng hạn của sinh viên.

# TIỀN XỬ LÝ DỮ LIỆU

## Thông tin về dữ liệu

* Bộ dữ liệu Education\_dataset\_V2 (UIT Dataset) được xây dựng từ thông tin thu thập từ sinh viên Trường Đại học Công nghệ Thông tin (UIT), thuộc Đại học Quốc gia TP.HCM, nhằm hỗ trợ nghiên cứu về hiệu suất và hành vi học tập của sinh viên. Dữ liệu bao gồm chi tiết kết quả học tập qua nhiều năm và học kỳ, tạo điều kiện cho việc phân tích các yếu tố ảnh hưởng đến kết quả học tập, đồng thời đề xuất các giải pháp nâng cao chất lượng giảng dạy.
* Bộ dữ liệu được thu thập trong khoảng thời gian từ năm 2013 đến năm 2016. Education\_dataset\_V2 cung cấp thông tin đa dạng về sinh viên UIT bao gồm: thông tin cá nhân, kết quả học tập, điểm rèn luyện, chi tiết các điểm thành phần, điểm trung bình theo học kỳ / theo toàn khóa, các chứng chỉ ngoại ngữ, xếp loại anh văn, điểm thi tốt nghiệp trung học phổ thông, các giấy tờ hành chính liên quan, xếp loại học vụ, thông tin bảo lưu, thông tin tốt nghiệp, thông tin miễn giảm học phí.

## Tiền xử lý dữ liệu

* 1. **Khám phá dữ liệu**
  2. **Làm sạch dữ liệu**
* Với mỗi tệp dữ liệu định dạng Excel từ 15 tệp dữ liệu của bộ dữ liệu ‘Education\_Dataset\_v2.zip’, lần lượt thực hiện các thao tác sau đây để làm sạch bộ dữ liệu.
* Kiểm tra và loại bỏ các cột không cần thiết: df.drop()
* Kiểm tra các giá trị thiếu: df.isnull().sum()
* Xử lý các giá trị thiếu
* Điền giá trị trung bình cho biến số: df.fillna()
* Điền giá trị phổ biến nhất cho biến phân loại: df.fillna()
* Chuẩn hóa các biến cần thiết: pd.to\_numeric()
* Kiểm tra dữ liệu sau khi làm sạch
* Lưu dữ liệu vào file mới
  1. **Chuyển đổi dữ liệu**
  2. **Ghi nhãn dữ liệu**
  3. **Chia tập dữ liệu**

## Kết hợp dữ liệu

# PHÂN TÍCH DỮ LIỆU

## Phân tích

## Phân tích các yếu tố ảnh hưởng khác

## Thảo luận kết quả phân tích

# XÂY DỰNG MÔ HÌNH ỨNG DỤNG

## Chuẩn bị dữ liệu

## Phương pháp đánh giá

* Sử dụng các độ đo:
* *Accuracy*
* *Precision*
* *F1-Score*
* *Recall*

1. **Phương pháp thực nghiệm**

**3.1. Support Vector Machines (SVM)**

**3.1.1. Khái niệm**

Support Vector Machines (SVM) là một thuật toán học máy giám sát được sử dụng cho cả phân loại và hồi quy. Ý tưởng chính của SVM là tìm một siêu phẳng (hyperplane) tốt nhất để phân tách dữ liệu thuộc các lớp khác nhau. Thuật toán hoạt động bằng cách tối ưu hóa một biên (margin) giữa các lớp dữ liệu, sử dụng các điểm dữ liệu được gọi là "vector hỗ trợ" (support vectors).

**3.1.2. Đặc trưng**

* **Tính trực quan**: SVM không trực quan vì các siêu phẳng và hàm hạt nhân thường khó giải thích trực tiếp. Chúng hoạt động trên không gian chiều cao, điều này làm cho việc hình dung trở nên phức tạp.
* **Phù hợp với loại dữ liệu nào**: SVM làm việc tốt với dữ liệu có biên phân tách rõ ràng, đặc biệt trong không gian cao chiều. Tuy nhiên, hiệu suất có thể giảm với dữ liệu có lớp chồng chéo.
* **Khả năng mở rộng và kết hợp**: Hạn chế với dữ liệu rất lớn do chi phí tính toán cao, nhưng có thể cải thiện bằng cách sử dụng SVM tuyến tính hoặc biến thể ngẫu nhiên hóa.
* **Độ nhạy với dữ liệu nhiễu**: SVM khá nhạy cảm với dữ liệu nhiễu và giá trị ngoại lai, nhưng điều này có thể được giảm bớt bằng cách điều chỉnh tham số C.
* **Tiêu chí phân tách phổ biến**: Sử dụng hàm hạt nhân (kernel functions) như tuyến tính, Gaussian RBF, và đa thức để xác định cách các điểm dữ liệu được ánh xạ vào không gian cao hơn.
* **Yêu cầu tiền xử lý dữ liệu**: Thường yêu cầu chuẩn hóa hoặc chuẩn chỉnh dữ liệu để đạt hiệu quả cao.

**3.2. Logistic Regression**

**3.2.1. Khái niệm**

Logistic Regression là một mô hình học máy giám sát sử dụng để phân loại nhị phân. Nó dựa trên hồi quy tuyến tính nhưng sử dụng hàm sigmoid để ánh xạ đầu ra vào khoảng từ 0 đến 1, biểu diễn xác suất một mẫu thuộc về một lớp nhất định.

**3.2.2. Đặc trưng**

* **Tính trực quan:** Logistic Regression rất dễ giải thích nhờ bản chất tuyến tính. Hệ số trọng số (coefficients) có thể được giải thích là ảnh hưởng của từng đặc trưng đối với xác suất của lớp đích.
* **Phù hợp với loại dữ liệu nào:** Thích hợp với dữ liệu mà các đặc trưng có mối quan hệ tuyến tính với nhãn (hoặc logit của nhãn).
* **Khả năng mở rộng và kết hợp:** Có thể mở rộng tốt với dữ liệu lớn, đặc biệt khi kết hợp với các phương pháp tối ưu hóa gradient stochastic. Cũng dễ tích hợp vào các hệ thống mô hình lớn hơn.
* **Độ nhạy với dữ liệu nhiễu:** Dễ bị ảnh hưởng bởi các giá trị ngoại lai, nhưng có thể được giảm thiểu bằng cách sử dụng các phương pháp như regularization (L1, L2).
* **Tiêu chí phân tách phổ biến**: Sử dụng hàm sigmoid để ánh xạ đầu ra thành xác suất từ 0 đến 1.
* **Tốc độ và hiệu suất:** Là một mô hình nhanh, hiệu quả cho các bài toán nhỏ và vừa.

**3.3. Decision Tree**

**3.3.1. Khái niệm**

Decision Tree là một thuật toán học máy giám sát hoạt động dựa trên việc chia nhỏ dữ liệu thành các nhóm con bằng cách sử dụng các tiêu chí phân tách (split criteria). Mỗi nút trong cây đại diện cho một đặc trưng, và các nhánh đại diện cho kết quả của tiêu chí phân tách đó.

**3.3.2. Đặc trưng**

* **Tính trực quan:** Cây quyết định cực kỳ trực quan, dễ hiểu và giải thích. Cấu trúc dạng cây giúp người dùng dễ dàng theo dõi và phân tích quá trình ra quyết định.
* **Phù hợp với loại dữ liệu nào**: Làm việc tốt với cả dữ liệu định lượng (số) và định tính (phân loại). Không yêu cầu tiền xử lý phức tạp như chuẩn hóa hoặc scaling.
* **Khả năng mở rộng và kết hợp:** Cây đơn lẻ có thể giới hạn về hiệu suất trên dữ liệu phức tạp, nhưng khi kết hợp thành các hệ thống như Random Forest hoặc Gradient Boosting, khả năng mở rộng được cải thiện đáng kể.
* **Độ nhạy với dữ liệu nhiễu:** Cây quyết định rất dễ bị overfitting nếu không giới hạn độ sâu hoặc số nút lá, dẫn đến nhạy cảm với dữ liệu nhiễu.
* **Tiêu chí phân tách phổ biến**: Sử dụng các tiêu chí như Gini index, Entropy (còn gọi là thông tin lợi ích).
* **Tính bền vững:** Với dữ liệu nhỏ và đơn giản, cây quyết định là một trong những thuật toán mạnh và hiệu quả.

**3.4. Random Forest**

**3.4.1. Khái niệm**

Random Forest là một thuật toán học máy sử dụng nhiều cây quyết định được huấn luyện trên các tập dữ liệu con (subsets) và kết hợp kết quả bằng cách bỏ phiếu (với bài toán phân loại) hoặc lấy trung bình (với bài toán hồi quy).

**3.4.2. Đặc trưng**

* **Tính trực quan**: Kém trực quan hơn so với một cây đơn lẻ, vì mô hình là sự tổng hợp của hàng trăm hoặc hàng nghìn cây quyết định. Tuy nhiên, việc phân tích tầm quan trọng của từng đặc trưng trong mô hình giúp tăng tính giải thích.
* **Phù hợp với loại dữ liệu nào:** Làm việc tốt với dữ liệu phức tạp, bao gồm cả dữ liệu có mối quan hệ phi tuyến. Hiệu quả với cả bài toán phân loại và hồi quy.
* **Khả năng mở rộng và kết hợp:** Khả năng mở rộng tốt với dữ liệu lớn và dễ dàng song song hóa do các cây được xây dựng độc lập.
* **Độ nhạy với dữ liệu nhiễu:** Random Forest giảm thiểu overfitting so với cây quyết định nhờ phương pháp tổng hợp (bagging), nhưng vẫn có thể bị ảnh hưởng bởi dữ liệu nhiễu trong một số trường hợp.
* **Tiêu chí phân tách phổ biến:** Gini index, Entropy, hoặc các tiêu chí tùy chỉnh cho bài toán cụ thể.
* **Tính ổn định:** Ổn định hơn cây quyết định, hiệu suất thường không bị giảm mạnh với các thay đổi nhỏ trong dữ liệu.

**3.5. LightGBM**

**3.5.1. Khái niệm**

LightGBM (Light Gradient Boosting Machine) là một thuật toán tăng cường độ dốc (gradient boosting) được tối ưu hóa cho tốc độ và hiệu suất trên các tập dữ liệu lớn. Nó xây dựng các cây dựa trên gradient với chiến lược tăng trưởng tập trung (leaf-wise growth).

**3.5.2. Đặc trưng**

* **Tính trực quan**: Ít trực quan hơn so với các mô hình đơn giản như Logistic Regression hoặc Decision Tree. Mô hình phức tạp và các chiến lược tăng cường khó giải thích trực tiếp, nhưng các công cụ như SHAP hoặc LIME có thể hỗ trợ.
* **Phù hợp với loại dữ liệu nào**: Tốt với dữ liệu lớn, đa chiều, và có nhiều đặc trưng không đồng nhất. Cũng hoạt động hiệu quả với dữ liệu thiếu thông tin (missing data).
* **Khả năng mở rộng và kết hợp**: Được tối ưu hóa để làm việc với dữ liệu lớn trên nhiều lõi xử lý hoặc hệ thống phân tán. Dễ dàng tích hợp vào các pipeline học máy.
* **Độ nhạy với dữ liệu nhiễu:** Nhạy cảm với nhiễu nếu không điều chỉnh tham số đúng cách (như learning rate, max depth).
* **Tiêu chí phân tách phổ biến**: Gradient-based splitting, tối ưu hóa dựa trên giá trị gradient của hàm mất mát.
* **Hiệu suất và tốc độ:** Nhanh hơn so với các mô hình Gradient Boosting khác (như XGBoost), nhờ vào việc sử dụng histogram-based binning và chiến lược tăng trưởng leaf-wise.
* **Tính linh hoạt:** Cho phép điều chỉnh linh hoạt nhiều tham số để tối ưu hóa hiệu suất cho từng bài toán cụ thể.

## Kết quả thực nghiệm

### Thực nghiệm cơ bản

### Thực nghiệm mở rộng

# KẾT LUẬN

# 

# TÀI LIỆU THAM KHẢO

1. Trường Đại học Công nghệ Thông tin. (n.d.). *Quyết định về việc ban hành Quy chế đào tạo theo học chế tín chỉ cho hệ đại học chính quy của Trường Đại học Công nghệ Thông tin*. Truy cập ngày 15/11/2024, từ<https://daa.uit.edu.vn/content/qui-che-qui-dinh-qui-trinh-dao-tao-dai-hoc>.