

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**  
**ĐẠI HỌC ĐÀ NẴNG**

**NGUYỄN THỊ ÁNH HỒNG**

**NGHIÊN CỨU CÁC KỸ THUẬT PHÂN LỚP  
DỮ LIỆU VÀ XÂY DỰNG CHƯƠNG TRÌNH  
HỖ TRỢ ĐÁNH GIÁ THÀNH TÍCH NHÂN VIÊN  
CHO CÔNG TY ESILICON VIỆT NAM**

**Chuyên ngành : Khoa học máy tính**  
**Mã số : 60.48.01**

**TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT**

**Đà Nẵng - Năm 2012**

Công trình được hoàn thành tại  
**ĐẠI HỌC ĐÀ NẴNG**

Người hướng dẫn khoa học: **TS. TRƯƠNG NGỌC CHÂU**

Phản biện 1 : **PGS.TSKH. TRẦN QUỐC CHIẾN**

Phản biện 2 : **TS. TRẦN THIÊN THÀNH**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 20 tháng 01 năm 2013

*Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng;
- Trung tâm Học liệu, Đại học Đà Nẵng.

## MỞ ĐẦU

### 1. LÝ DO CHỌN ĐỀ TÀI

Nhiều doanh nghiệp đã phải hứng chịu hậu quả vì sự thiếu hụt việc xây dựng và duy trì hệ thống đánh giá và phát triển nguồn nhân lực một cách chặt chẽ và khoa học. Có nơi thì cán bộ chủ chốt bất mãn nghỉ việc, nhân viên cấp dưới hoang mang. Nơi khác thì mất khách hàng do chất lượng dịch vụ giảm sút, sản phẩm bị lỗi nhiều làm tăng chi phí. Chính vì những lý do trên, tôi quyết định chọn đề tài **“Nghiên cứu các kỹ thuật phân lớp dữ liệu và xây dựng chương trình hỗ trợ đánh giá thành tích nhân viên cho công ty eSilicon”** nhằm hỗ trợ cho phòng nhân sự, ban giám đốc cũng như các quản lý nhóm có được cái nhìn tổng quan, đánh giá được bao quát hiệu quả làm việc của nhân viên.

### 2. NHIỆM VỤ NGHIÊN CỨU

Mục tiêu mà đề tài hướng đến là nghiên cứu các kỹ thuật phân lớp dữ liệu và áp dụng kỹ thuật phân lớp C4.5 vào việc xây dựng chương trình hỗ trợ đánh giá thành tích nhân viên tại công ty eSilicon Việt Nam.

### 3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng và phạm vi nghiên cứu của đề tài cụ thể như sau:

- Đối tượng nghiên cứu:
  - o Nhân viên đang làm việc tại công ty.
  - o Quy trình và kết quả đạt được của nhân viên qua các dự án.
  - o Quy trình và các phương pháp phân lớp, kỹ thuật hỗ trợ ra quyết định trong khai phá dữ liệu.
- Phạm vi nghiên cứu:

- Số liệu thống kê về tình hình thực thi công việc được giao của nhân viên qua các dự án và các quyết định tăng lương của các nhân viên.
- Nghiên cứu lý thuyết trợ giúp ra quyết định, xây dựng ứng dụng hỗ trợ đánh giá thành tích cho nhân viên.

#### **4. PHƯƠNG PHÁP NGHIÊN CỨU**

- Về mặt lý thuyết:
  - Nghiên cứu tài liệu và các công nghệ liên quan.
  - Tổng hợp, thu thập các tài liệu về việc ghi nhận đánh giá thành tích nhân viên qua các dự án.
- Về mặt thực nghiệm:
  - Phân tích yêu cầu thực tế của bài toán và áp dụng các thuật toán liên quan để trợ giúp cho việc đưa ra các đánh giá.
  - Thống kê, phân tích các số liệu thực tế trong quy trình đánh giá
  - Kiểm tra, thử nghiệm và đưa ra các đánh giá kết quả đạt được.

#### **5. DỰ KIẾN KẾT QUẢ ĐẠT ĐƯỢC**

- Xây dựng kho dữ liệu lưu trữ thành tích làm việc của nhân viên.
- Đề xuất giải pháp kỹ thuật để đánh giá thành tích nhân viên trong công ty eSilicon nhanh chóng, có độ chính xác cao và có giá trị đối với nhà quản lý và điều hành.

#### **6. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN**

- Về mặt khoa học:
  - Đề tài đưa ra một phương thức ứng dụng cây quyết định trong công tác đánh giá nhân viên, tạo tiền đề cho các ứng dụng quản lý sau này.

- Về mặt thực tiễn:
  - Ứng dụng phần mềm mã nguồn mở weka để hỗ trợ đưa ra các đánh giá thành tích nhân viên có độ chuẩn xác và tính công bằng cao.
  - Triển khai và xây dựng mở rộng ứng dụng tại công ty eSilicon Việt Nam nhằm phục vụ tốt hơn công tác đánh giá nhân viên.

## **7. BỐ CỤC LUẬN VĂN**

Nội dung chính của luận văn được chia thành 3 chương như sau:

- Chương 1: Trình bày cơ sở lý thuyết làm nền tảng, bao gồm: tổng quan về khai phá dữ liệu, các phương pháp phân lớp, cây quyết định và hệ trợ giúp ra quyết định.
- Chương 2: Tìm hiểu, giới thiệu và phân tích thực trạng công tác đánh giá nhân viên, những hạn chế và giải pháp khắc phục, áp dụng giải pháp ứng dụng cây quyết định vào việc hỗ trợ đánh giá như thế nào.
- Chương 3: Trình bày chi tiết mô hình kho dữ liệu, cấu trúc tổng thể của hệ thống hỗ trợ đánh giá. Thống kê và đánh giá kết quả thử nghiệm.

## CHƯƠNG 1

### TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

#### 1.1 KHAI PHÁ DỮ LIỆU - DATA MINING LÀ GÌ?

Data Mining là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó. Data Mining được chia nhỏ thành một số hướng nghiên cứu chính như: mô tả khái niệm, luật kết hợp, phân lớp và dự đoán, phân cụm, khai phá chuỗi.

#### 1.2 GIỚI THIỆU VỀ PHÂN LỚP TRONG DATA MINING

*Phân lớp* và *dự đoán* là hai dạng của phân lớp dữ liệu. *Phân lớp* dự đoán giá trị của những nhãn xác định hay những giá trị rời rạc. *Dự đoán* xây dựng mô hình với các hàm nhận giá trị liên tục.

##### 1.2.1 Các bước của quá trình phân lớp

Quá trình phân lớp dữ liệu gồm hai bước như sau: Quá trình học nhằm mô tả một tập các lớp dữ liệu và quá trình phân lớp các mô tả này.

##### 1.2.2 Các vấn đề liên quan đến phân lớp dữ liệu

###### 1.2.2.1 Chuẩn bị dữ liệu cho việc phân lớp

Chuẩn bị dữ liệu hay còn gọi là quá trình tiền xử lý dữ liệu, gồm có việc làm sạch dữ liệu và chuyển đổi dữ liệu. Quá trình này giúp cải thiện độ chính xác, tính hiệu quả và khả năng mở rộng được của mô hình phân lớp.

###### 1.2.2.2 Các phương pháp đánh giá độ chính xác của mô hình phân lớp

Có 2 phương pháp đánh giá phổ biến là holdout và k-fold cross-validation.

## 1.3 PHƯƠNG PHÁP PHÂN LỚP VỚI CÂY QUYẾT ĐỊNH

### 1.3.1 Cây quyết định và luật

### 1.3.2 Ưu điểm của phương pháp khai phá dữ liệu bằng cây QĐ

Cây quyết định là phương pháp có một số ưu điểm như dễ hiểu, không đòi hỏi việc chuẩn hóa dữ liệu, có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn...

### 1.3.3 Giới thiệu thuật toán C4.5

Giải thuật C4.5 biểu diễn các khái niệm ở dạng các cây quyết định. Giải thuật có đầu vào, đầu ra như sau:

Đầu vào: Tập dữ liệu huấn luyện - là tập hợp các ví dụ.

Đầu ra: Cây quyết định tương ứng với tập dữ liệu huấn luyện đó.

#### \* Thuộc tính nào là thuộc tính dùng để phân loại tốt nhất?

Năm 1948, Shannon đưa ra lý thuyết thông tin cung cấp một khái niệm để đo tính thuần nhất của một tập hợp gọi là entropy. Giả sử các ví dụ của tập S thuộc i loại và có C giá trị phân loại thì công thức entropy tổng quát là:

$$\text{Entropy}(S) = - \sum_{i=1}^c p_i \log_2 p_i$$

Với C : số giá trị phân loại

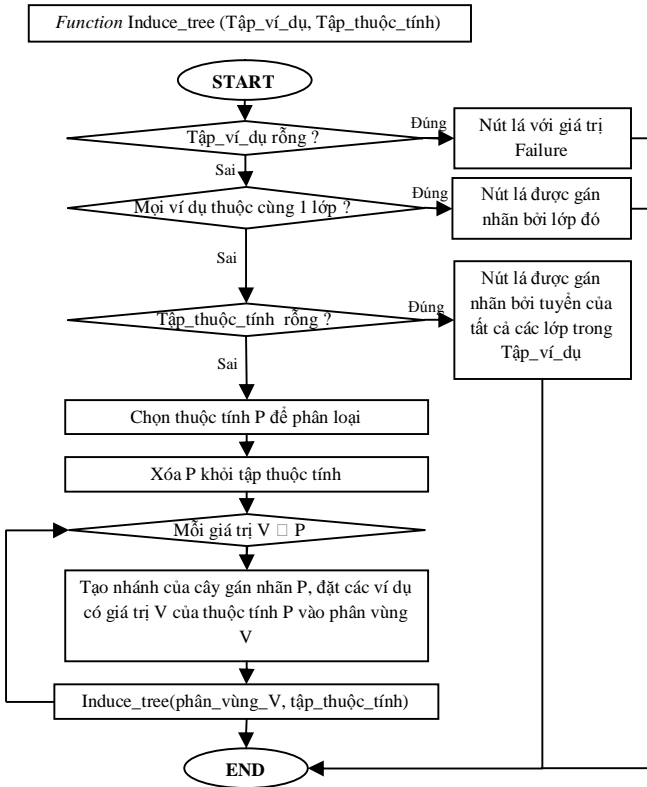
#### \* Độ lợi thông tin - Information Gain

Độ lợi thông tin - *Information gain*, là một phép đo hiệu suất phân loại các ví dụ của một thuộc tính. Ví dụ,  $\text{Gain}(S, A)$  của thuộc tính A, trên tập S, được định nghĩa như sau:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Trong đó,  $\text{Values}(A)$  là tập hợp có thể có các giá trị của thuộc tính A, và  $S_v$  là tập con của S chứa các ví dụ có thuộc tính A mang giá trị v.

**Thuật toán: C4.5 xây dựng cây quyết định theo giải thuật sau:**



Hình 1.9 Sơ đồ khối của quá trình xây dựng cây quyết định trong thuật toán C4.5

### 1.3.4 Tỷ suất lợi ích GainRatio

Thuật toán C4.5 mở rộng cách tính Information Gain thành Gain Ratio để cố gắng khắc phục sự thiên lệch. Gain Ratio được xác định bởi công thức sau:

$$\text{GainRatio}(S,P) = \frac{\text{Gain}(S,P)}{\text{SplitInfo}(S,P)}$$



Với  $\text{SplitInfo}(S,P)$  chính là thông tin do phân tách  $P$  trên cơ sở giá trị của thuộc tính phân loại  $S$ . Công thức tính như sau:

$$\text{SplitInfo}(S,P) = - \sum_{i=1}^C \frac{|P_i|}{|P|} \text{Log} \frac{|P_i|}{|P|}$$

*Trong đó:*

- +  $P$  là tập các giá trị thuộc tính của  $S$
- +  $P_i$  là tập con của tập  $P$  ứng với thuộc tính  $S$  giá trị  $v_i$
- +  $C$  là số giá trị phân loại

### **1.3.5 Cắt xén cây quyết định và Suy luận tập hợp luật**

#### ***1.3.5.1 Cắt xén cây quyết định***

Cắt xén cây quyết định được thực hiện bởi việc thay thế toàn bộ một nhánh con bởi một nút lá. Sự thay thế này xảy ra nếu một luật quyết định chứng minh rằng tỷ lệ lỗi mong đợi trong cây con lớn hơn trong 1 nút lá.

#### ***1.3.5.2 Suy luận tập hợp luật***

Để dễ dàng suy luận tập hợp luật, chúng ta viết tương ứng mỗi luật cho một nhánh trong cây quyết định từ gốc cho đến một nút lá. Trong luật đó, phía bên trái được xây dựng dễ dàng từ nhãn của các nút lá và nhãn của các cung.

## **TỔNG KẾT CHƯƠNG 1**

Trong chương 1 này, luận văn đã trình bày khái quát về khái niệm khai phá dữ liệu, phân lớp và các vấn đề liên quan đến phân lớp dữ liệu.

Trong chương tiếp theo, luận văn sẽ đi sâu vào phân tích thực trạng đánh giá thành tích nhân viên tại công ty eSILICON Việt Nam, để từ đó có thể hiểu rõ thực trạng hệ thống hiện tại và đưa ra các giải pháp phù hợp nhằm xây dựng hệ thống trợ giúp ra quyết định trong việc đánh giá thành tích nhân viên.

## CHƯƠNG 2

# PHÂN TÍCH THỰC TRẠNG CÔNG TÁC ĐÁNH GIÁ THÀNH TÍCH NHÂN VIÊN TẠI CÔNG TY ESILICON

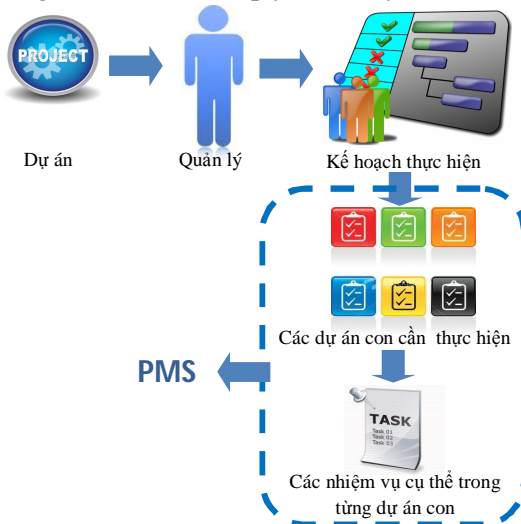
### 2.1 GIỚI THIỆU CÔNG TY ESILICON VIỆT NAM

eSilicon Việt Nam xuất thân từ công ty cổ phần SDS-Silicon Design Solutions Việt Nam, một trong những công ty chuyên thiết kế bộ nhớ nhúng (Embedded Memory) và cung cấp các dịch vụ thiết kế vi mạch tích hợp (ASIC) trên các công nghệ đúc và thiết bị bán dẫn.

### 2.2 CÔNG TÁC ĐÁNH GIÁ THÀNH TÍCH NHÂN VIÊN

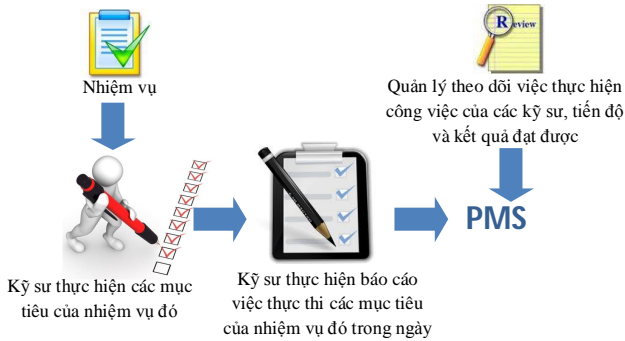
#### 2.2.1 Giới thiệu hệ thống quản lý công việc PMS

Quy trình quản lý của hệ thống PMS được bắt đầu từ khâu các quản lý nhận một dự án và lên kế hoạch thực hiện dự án đó với lượng kỹ sư và thời gian xác định. Và quy trình này có thể được mô tả:



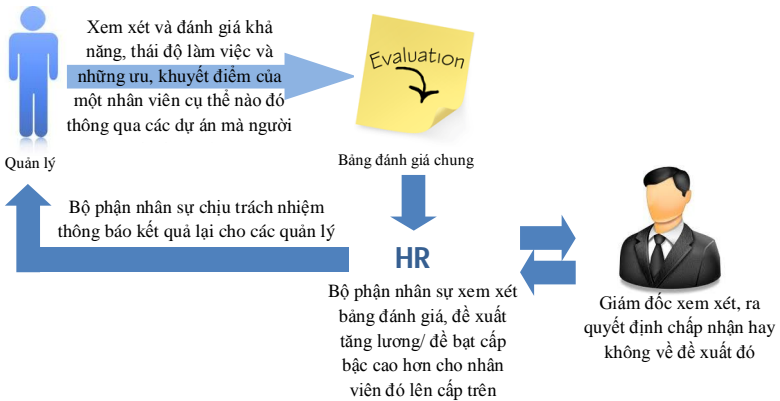
Hình 2.1 - Quy trình quản lý dự án của hệ thống PMS

Đối với các kỹ sư, quy trình quản lý bắt đầu từ việc thực hiện mỗi nhiệm vụ được giao, báo cáo việc thực hiện công việc đó như thế nào trong mỗi ngày về công việc được giao.



Hình 2.2 - Quy trình quản lý chung

## 2.2.2 Quy trình đánh giá thành tích nhân viên



Hình 2.3 - Quy trình đánh giá thành tích nhân viên

### 2.2.3 Các chỉ tiêu đánh giá thành tích nhân viên

Sau mỗi một dự án, nhân viên được đánh giá dựa vào một số các tiêu chí và tương ứng với mỗi tiêu chí có 5 mức ứng với 5 thang điểm từ 1 đến 5 để đánh giá. Cụ thể gồm các tiêu chí như sau:

#### 2.2.3.1 Khả năng lãnh đạo

#### 2.2.3.2 Khả năng giao tiếp

#### 2.2.3.3 Khả năng quản lý

#### 2.2.3.4 Khả năng giao công việc

#### 2.2.3.5 Khả năng làm việc

#### 2.2.3.6 Khả năng quản lý dự án

#### 2.2.3.7 Khả năng phân tích và giải quyết vấn đề

### 2.2.4 Mối quan hệ tương quan giữa các tiêu chí và việc tăng lương

Đối với bài toán luận văn đặt ra, chúng tôi dựa vào số liệu thống kê tăng lương của các nhân viên tùy thuộc vào chức vụ, kết quả đánh giá chung và các tiêu chí cần cho mỗi chức vụ khác nhau. Trong đó, kết quả đánh giá chung được tính toán dựa trên giá trị trung bình của các tiêu chí:

$$R = \frac{\sum \text{LDS} + \sum \text{CM} + \sum \text{MP} + \sum \text{DLG} + \sum \text{WP} + \sum \text{PM} + \sum \text{PIR}}{7 \cdot n}$$

Với:

- n – số lượng đánh giá trong thời gian xác định
- LDS – Leadership: khả năng lãnh đạo
- CM – Communication : khả năng giao tiếp
- MP – Managing Performance : Khả năng quản lý
- DLG – Delegation : Khả năng giao công việc
- WP – Working Performance : Khả năng làm việc

- PM – Project Management : Khả năng quản lý dự án
- PIR - Problem Identification and Resolution: Khả năng phân tích và giải quyết vấn đề
- SI – Salary Increase – Kết quả tăng lương gần nhất
- Eff. – Effort – Hiệu quả của việc tăng lương so với mức đáp ứng trong công việc.

*Bảng 2.1 – Bảng dữ liệu hỗ trợ đánh giá tăng lương*

Chức vụ  Tiêu chí Đánh giá	Manager			Leader			Employee		
	Kết quả đánh giá chung			Kết quả đánh giá chung			Kết quả đánh giá chung		
	$1 \leq R < 2$	$2 \leq R < 3$	$3 \leq R \leq 5$	$1 \leq R < 2$	$2 \leq R < 3$	$3 \leq R \leq 5$	$1 \leq R < 2$	$2 \leq R < 3$	$3 \leq R \leq 5$
$1 \leq LDS < 2$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$2 \leq LDS < 3$	NO	NO	NO	NO	NO	YES	NO	YES	YES
$3 \leq LDS < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq CM < 2$	NO	NO	NO	NO	NO	NO	NO	NO	NO
$2 \leq CM < 3$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$3 \leq CM < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq MP < 2$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$2 \leq MP < 3$	NO	NO	NO	NO	NO	YES	NO	YES	YES
$3 \leq MP < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq DLG < 2$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$2 \leq DLG < 3$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$3 \leq DLG < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq WP < 2$	NO	NO	NO	NO	NO	NO	NO	NO	NO
$2 \leq WP < 3$	NO	NO	NO	NO	NO	NO	NO	NO	NO
$3 \leq WP < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq PM < 2$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$2 \leq PM < 3$	NO	NO	NO	NO	NO	YES	NO	YES	YES
$3 \leq PM < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES
$1 \leq PIR < 2$	NO	NO	NO	NO	NO	NO	NO	NO	NO
$2 \leq PIR < 3$	NO	NO	NO	NO	NO	NO	NO	YES	YES
$3 \leq PIR < 5$	NO	NO	YES	NO	NO	YES	NO	YES	YES

Trong đó:

- ‘YES’ trong bảng trên tương ứng với việc nhân viên đó đáp ứng được toàn bộ các tiêu chí đặt ra tương ứng với chức vụ và được tăng lương.
- ‘NO’ trong bảng trên tương ứng với việc nhân viên đó không đáp ứng được toàn bộ các tiêu chí đặt ra tương ứng với chức vụ và không được tăng lương.

### **2.2.5 Vấn đề trợ giúp quyết định**

Các quản lý thường đánh giá nhân viên thông qua các công việc được báo cáo trên hệ thống PMS qua một khoảng thời gian xác định (thông thường là 6 tháng hoặc 1), việc đánh giá này thường có thể ẩn chứa những nguy cơ mang các yếu tố cảm tính, không thực sự chính xác. Việc xây dựng hệ thống hỗ trợ đánh giá thành tích nhân viên sẽ góp phần gia tăng độ chuẩn xác, công bằng trong các đánh giá. Từ đó, giúp các nhà quản lý đưa ra các chế độ đãi ngộ phù hợp để giữ chân người tài.

### **2.2.6 Giải pháp xây dựng hệ thống trợ giúp quyết định**

Với một khối lượng dữ liệu lớn của trên 200 nhân viên qua hơn 10 năm thành lập về các vấn đề liên quan đến các đánh giá thành tích của các nhân viên và các quyết định tăng lương, thì yêu cầu đặt ra là cần phải có một phương pháp phân tích dữ liệu khoa học, nhằm đánh giá một cách chính xác các quyết định tăng lương đã qua so với mức độ đáp ứng công việc hiện tại của nhân viên.

Và giải pháp sử dụng cây quyết định trong thuật toán C4.5 là một giải pháp phù hợp để xây dựng hệ trợ giúp quyết định trong công tác đánh giá việc tăng lương cho các nhân viên thông qua các thành tích công việc của họ. Giải pháp này có thể xử lý được khối lượng dữ liệu lớn với tốc độ tính toán nhanh và có độ chính xác cao.

## TỔNG KẾT CHƯƠNG 2

Trong chương 2, luận văn đã trình bày các quy trình đánh giá cũng như hệ thống lưu trữ công việc hiện tại của công ty eSilicon Việt Nam. Bên cạnh đó, chương này cũng đã làm rõ việc đánh giá kết quả tăng lương của nhân viên thông qua các tiêu chí đánh giá khả năng đáp ứng trong công việc của họ.

Vấn đề được đặt ra trong chương này là liệu quyết định đánh giá tăng lương đã qua và mức độ đáp ứng trong công việc liệu có tương xứng hay không?

Vì vậy, trong chương tiếp theo, luận văn sẽ trình bày cách áp dụng cây quyết định vào việc xây dựng hệ thống trợ giúp ra quyết định đánh giá thành tích nhân viên tương ứng với các quyết định tăng lương của họ.



## **CHƯƠNG 3**

### **ỨNG DỤNG CÂY QUYẾT ĐỊNH VÀO XÂY DỰNG HỆTHỐNG**

#### **3.1 PHÂN TÍCH THỰC TRẠNG VÀ CHỨC NĂNG CỦA HỆ THỐNG**

##### **3.1.1 Chức năng của hệ thống**

Các chức năng chính của hệ thống:

Lưu trữ đánh giá nhân viên

Thống kê báo cáo

Hỗ trợ ra quyết định trong việc đánh giá tăng lương

##### **3.1.2 Phân tích các yêu cầu của hệ thống**

Yêu cầu của người sử dụng: Lưu trữ được các đánh giá của nhân viên, thống kê báo cáo các đánh giá và đưa ra kết quả đánh giá tổng quan, hỗ trợ ra quyết định.

Yêu cầu của người phát triển hệ thống

- Yêu cầu về công nghệ: Hiểu biết về nguồn dữ liệu, vấn đề làm mịn và tái tạo dữ liệu, xử lý và phân tích dữ liệu.
- Yêu cầu về triển khai: Lập các báo cáo thống kê hỗ trợ ra quyết định, phân tích dữ liệu nhiều chiều và dữ liệu thống kê, cho phép truy nhập và truy cập các thông tin cần thiết.

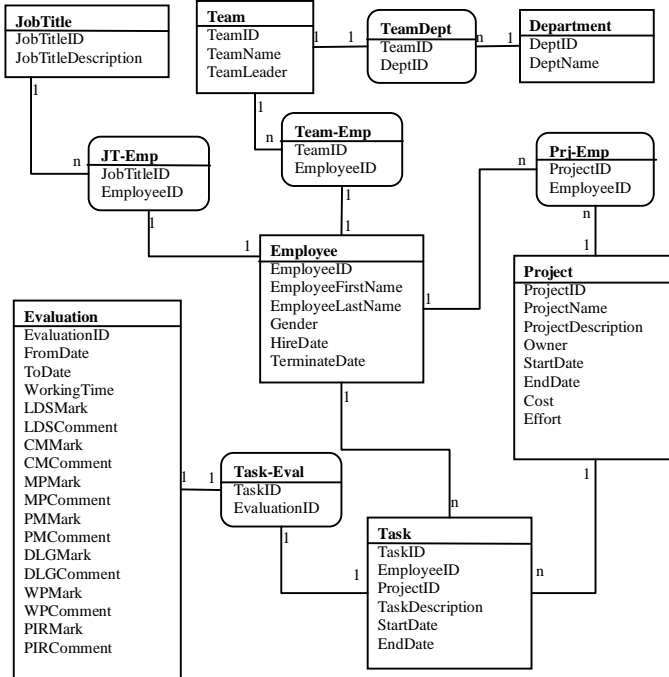
##### **3.1.3 Xây dựng cơ sở dữ liệu tác nghiệp**

Quy trình nhập đánh giá

Quy trình thống kê báo cáo

Quy trình hỗ trợ ra quyết định

Mô hình ERD mô tả mối quan hệ giữa các bảng và thực thể:



Hình 3.4 – Mô hình ERD

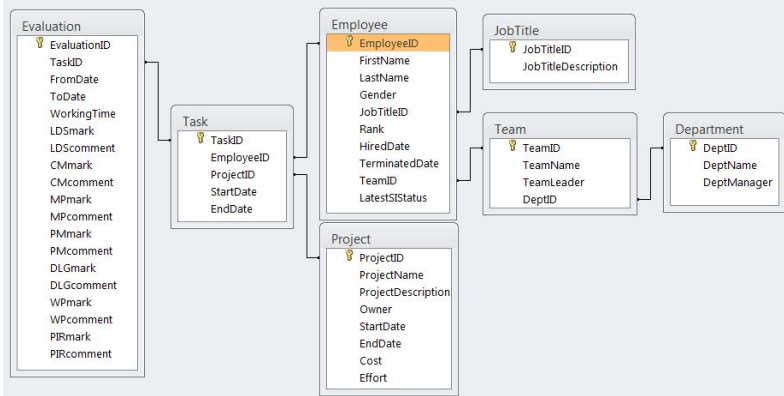
Mô hình dữ liệu đa chiều

Kho dữ liệu được thiết kế gồm 2 mảng chính:

- Thống kê đánh giá của nhân viên
- Hỗ trợ ra quyết định đánh giá thành tích nhân viên

### 3.1.4 Thiết kế kho dữ liệu

#### 3.1.4.1 Tổ chức mô hình kho dữ liệu



Hình 3.6– Kho dữ liệu

#### 3.2.4.2 Các chiều và bảng sự kiện

- Bảng JobTitle – Đặc tả loại hình công việc
- Bảng Department – Thông tin phòng ban
- Bảng Team – Thông tin đội nhóm làm việc
- Bảng Employee – Thông tin nhân viên
- Bảng Task – Thông tin việc sắp xếp/bố trí công việc
- Bảng Project – Thông tin dự án
- Bảng Evaluation – Thông tin đánh giá nhân viên

### 3.1.5 Phân tích dữ liệu

Qua phân tích thực trạng và các chức năng của hệ thống, chúng ta thấy rằng có nhiều yếu tố ảnh hưởng đến việc đưa ra quyết định tăng lương cho một nhân viên, bao gồm: Chức vụ, kỹ năng cần tương ứng với các chức vụ và tỷ lệ tăng lương của các phòng ban.

Các yếu tố chính này tạo nên các tập thuộc tính để dự đoán giá trị cho quyết định tăng lương và tỷ lệ tăng tương ứng.

Từ những số liệu đã có, ta có thể đưa ra bảng giá định về đánh giá của các nhân viên như sau:

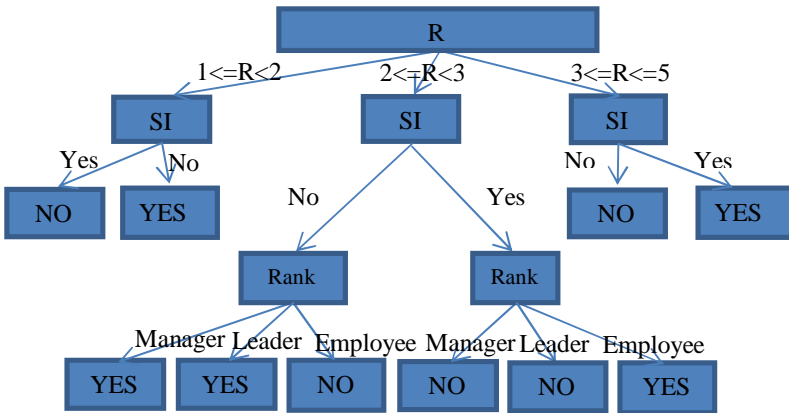
*Bảng 3.8 Tập dữ liệu huấn luyện để xây dựng cây quyết định cho bài toán hỗ trợ ra quyết định tăng lương*

Emp.	Rank	Dept.	LDS	CM	MP	DLG	WP	PM	PIR	R	SI	EFF.
E1	Manager	IPQA	5	3	5	5	5	3	3	<b>4.143</b>	Yes	Yes
E2	Employee	IC	1	2	1	1	2	1	3	<b>1.571</b>	Yes	No
E3	Leader	IPDE	5	4	5	5	5	3	3	<b>4.286</b>	No	No
E4	Leader	HR	3	3	1	3	2	3	2	<b>2.428</b>	No	Yes
E5	Employee	HR	1	2	1	1	2	1	3	<b>1.571</b>	Yes	No
E6	Employee	HR	2	4	3	2	3	1	5	<b>2.857</b>	Yes	Yes
E7	Leader	IPQA	2	3	3	2	3	3	2	<b>2.429</b>	No	Yes
E8	Leader	IPDE	5	4	3	5	5	5	3	<b>4.286</b>	Yes	Yes
E9	Manager	ASIC	2	3	5	2	3	2	2	<b>2.714</b>	Yes	No
E10	Employee	IPQA	3	3	3	3	3	1	5	<b>3</b>	No	No
E11	Leader	IC	3	4	5	3	3	2	3	<b>3.143</b>	Yes	Yes
E12	Employee	ASIC	3	4	1	3	3	1	3	<b>2.571</b>	Yes	Yes
E13	Employee	IPDE	3	3	3	2	3	1	3	<b>2.571</b>	No	No
E14	Employee	ASIC	2	3	3	2	5	1	5	<b>3</b>	Yes	Yes
E15	Manager	IC	5	3	3	3	5	3	5	<b>3.857</b>	No	No
E16	Leader	Layout	2	2	5	2	3	3	2	<b>2.714</b>	Yes	No
E17	Manager	Layout	3	4	5	3	3	3	3	<b>3.428</b>	Yes	Yes
E18	Manager	Layout	2	3	2	1	3	2	2	<b>2.142</b>	No	Yes
E19	Employee	IPDE	1	2	1	1	2	1	2	<b>1.428</b>	No	Yes
E20	Employee	IC	3	3	3	3	5	2	3	<b>3.142</b>	No	No
E21	Leader	ASIC	4	3	3	4	5	2	4	<b>3.571</b>	No	No
E22	Manager	IPDE	5	4	5	5	5	3	5	<b>4.571</b>	No	No
E23	Employee	IPQA	3	5	5	5	5	3	5	<b>4.428</b>	Yes	Yes
E24	Manager	IC	5	3	5	4	4	4	4	<b>4.142</b>	No	No
E25	Employee	IC	3	5	3	4	5	3	5	<b>4</b>	No	No

Từ đó, áp dụng thuật toán C4.5 cho tập dữ liệu huấn luyện để xây dựng cây quyết định cho bài toán hỗ trợ ra quyết định tăng lương.

### 3.2 TRIỂN KHAI GIẢI THUẬT C4.5 XÂY DỰNG CÂY QUYẾT ĐỊNH

Để xây dựng cây quyết định ta phải xác định nút gốc để phân tách cây. Thuộc tính có độ lợi thông tin lớn nhất sẽ được chọn làm nút gốc. Ứng với mỗi nhánh, ta lại tiếp tục vận dụng thuật toán để tìm thuộc tính tiếp theo làm nút của cây. Cuối cùng, ta có được cây quyết định hoàn chỉnh như trong hình 3.13:



Hình 3.13 – Cây quyết định hoàn chỉnh

### 3.3 XÂY DỰNG TẬP LUẬT TỪ CÂY QUYẾT ĐỊNH

Trong cây quyết định, mỗi một đường dẫn từ gốc đến nút lá tạo thành một luật, luật này có vẻ trái là một bộ giá trị của các thuộc tính được chọn để phân lớp, vẻ phải là một trong các giá trị thuộc tính kết quả. Từ cây quyết định trên, ta có thể rút ra một số các tập luật sau:

1. **IF**  $(R < 2) \wedge (SI = YES)$  **THEN** EFF = NO
2. **IF**  $(R < 2) \wedge (SI = NO)$  **THEN** EFF = YES
3. **IF**  $(R \geq 3) \wedge (SI = YES)$  **THEN** EFF = YES
4. **IF**  $(R \geq 3) \wedge (SI = NO)$  **THEN** EFF = NO
5. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (RANK = EMPLOYEE) \wedge (SI = NO)$

**THEN EFF = NO**

6. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{EMPLOYEE}) \wedge (\text{SI} = \text{YES})$

**THEN EFF = YES**

7. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{LEADER}) \wedge (\text{SI} = \text{NO})$

**THEN EFF = YES**

8. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{LEADER}) \wedge (\text{SI} = \text{YES})$

**THEN EFF = NO**

9. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{MANAGER}) \wedge (\text{SI} = \text{NO})$

**THEN EFF = YES**

10. **IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{MANAGER}) \wedge (\text{SI} = \text{YES})$

**THEN EFF = NO**

Ví dụ, đối với luật 1 '**IF**  $(R < 2) \wedge (\text{SI} = \text{YES})$  **THEN EFF = NO**' có nghĩa là '*Nếu R nhỏ hơn 2 và SI bằng YES thì hiệu quả tăng lương là không có.*'

Hay luật 5 '**IF**  $(R \geq 2) \wedge (R < 3) \wedge (\text{RANK} = \text{EMPLOYEE}) \wedge (\text{SI} = \text{YES})$  **THEN EFF = YES**' có nghĩa là '*Nếu R lớn hơn bằng 2 và nhỏ hơn 3 đối với nhân viên có Rank = Employee và SI bằng YES thì hiệu quả tăng lương là có.*'

## KẾT LUẬN

### 1. Kết quả đạt được

Về mặt khoa học

- Luận văn đã tiến hành phân tích, tìm hiểu được quy trình đánh giá và tăng lương của công ty. Phát hiện ra những hạn chế cũng như những nguy cơ tiềm ẩn trong công tác đánh giá và lưu trữ đánh giá một cách chuẩn xác nhằm nâng cao hiệu quả của việc đánh giá và đưa ra các đề xuất tăng lương.
- Nắm được các phương pháp và các mô hình toán học, áp dụng để giải quyết yêu cầu đặt ra.
- Nghiên cứu và vận dụng giải thuật C4.5 để xây dựng mô hình dự đoán.

Về mặt thực tiễn

- Luận văn đã nêu được hướng giải quyết cũng như giải pháp kỹ thuật để xây dựng hệ thống hỗ trợ đánh giá nhân viên và quyết định tăng lương phù hợp.
- Việc kết hợp lý thuyết hệ trợ giúp ra quyết định và phân tích dữ liệu bằng cây quyết định giúp giảm thiểu đáng kể những sai sót trong quá trình đánh giá và đưa ra quyết định.

### 2. Hướng phát triển

Cần thử nghiệm hệ thống với khối lượng dữ liệu lớn để đánh giá lại độ tin cậy của cây quyết định hỗ trợ tăng lương.

Phối hợp các phòng Nhân sự để xây dựng và hoàn thiện tập cơ sở dữ liệu huấn luyện nhằm đưa hệ thống vào sử dụng một cách có hiệu quả.

Nghiên cứu vận dụng thuật toán C5.0, một cải tiến của C4.5, để giảm thiểu tỉ lệ lỗi, nâng cao hiệu suất và tăng cường độ tối ưu cho ứng dụng.

Tiếp tục phát triển, nâng cấp hệ thống, hỗ trợ kết nối, truy xuất và xử lý dữ liệu trực tiếp với hệ quản trị cơ sở SQL Server đồng thời mở rộng sang các hệ quản trị CSDL khác.