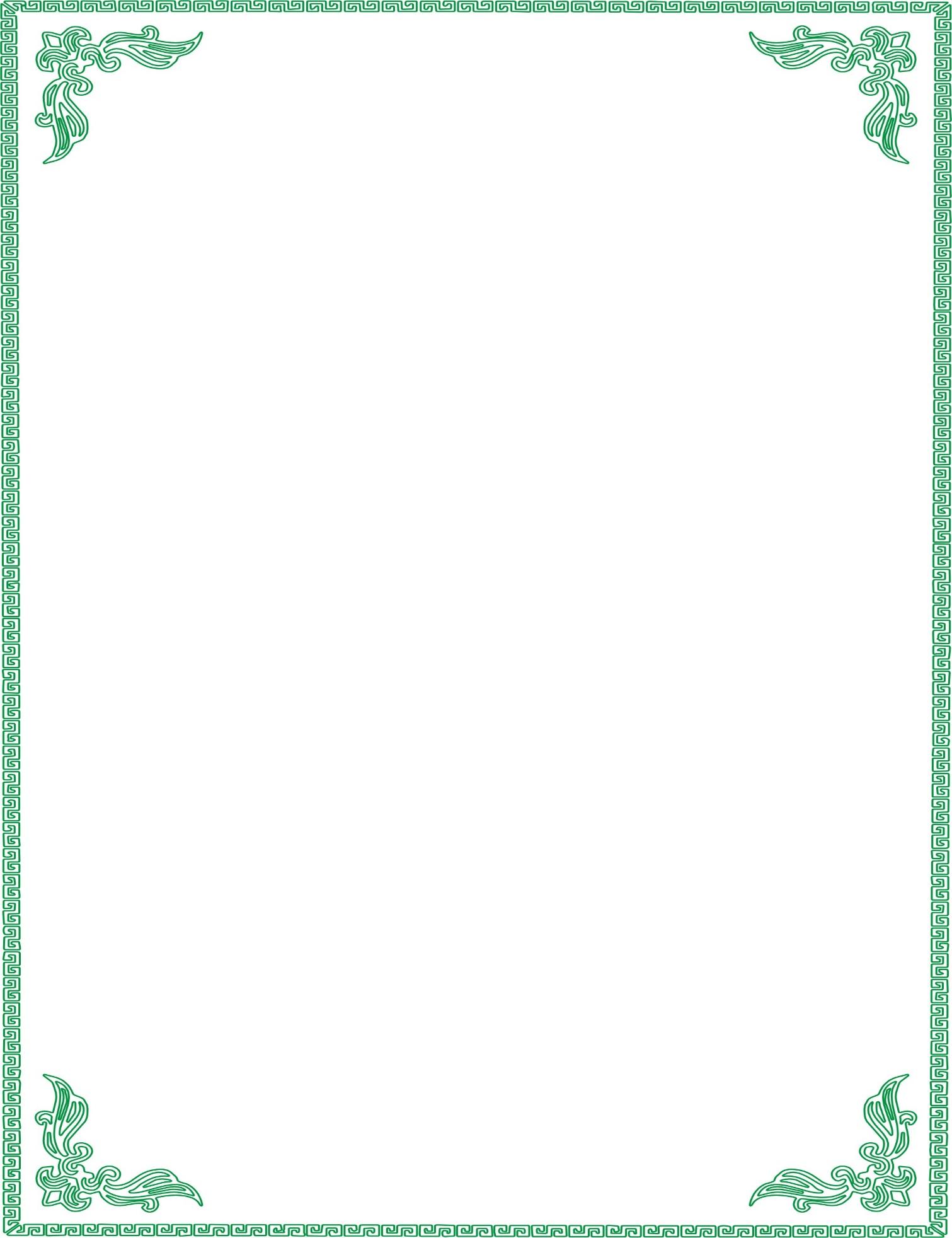
**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

 KHOA CÔNG NGHỆ THÔNG TIN

**BÀI TẬP LỚN MÔN DỮ LIỆU LỚN**

**ĐỀ 1: PHÂN TÍCH VÀ DỰ ĐOÁN CÁC NƯỚC**

**ĐI TRONG CỜ VUA**

**Giảng viên hướng dẫn: Ts. Trần Qúy Nam**

**ThS. Lê Thị Thùy Trang**

**Nhóm sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stt** | **Mã sinh viên** | **Họ và tên** | **Lớp** |
| **1** | **1671020292** | **Đinh Thế Thành** | **CNTT16-01** |
| **2** | **1671020072** | **ĐặngTrường Dương** | **CNTT16-01** |
| **3** | **1671020087** | **Trịnh Hoàng Hà** | **CNTT16-01** |
| **4** | **1671020278** | **Nguyễn Thanh Sơn** | **CNTT16-01** |

**Hà Nội, tháng 2 năm 2025**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

** KHOA CÔNG NGHỆ THÔNG TIN**

**BÀI TẬP LỚN MÔN DỮ LIỆU LỚN**

**ĐỀ 1: PHÂN TÍCH VÀ DỰ ĐOÁN CÁC NƯỚC**

**ĐI TRONG CỜ VUA**

**Giảng viên hướng dẫn: Ts. Trần Qúy Nam**

**ThS. Lê Thị Thùy Trang**

**Nhóm sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stt** | **Mã sinh viên** | **Họ và tên** | **Lớp** |
| **1** | **1671020292** | **Đinh Thế Thành** | **CNTT16-01** |
| **2** | **1671020072** | **Đặng Trường Dương** | **CNTT16-01** |
| **3** | **1671020087** | **Trịnh Hoàng Hà** | **CNTT16-01** |
| **4** | **1671020278** | **Nguyễn Thanh Sơn** | **CNTT16-01** |

**CÁN BỘ CHẤM THI 1 CÁN BỘ CHẤM THI 2**

**Hà Nội, tháng 2 năm 2025**

[**LỜI NÓI ĐẦU** 6](#_Toc192971582)

[**CHƯƠNG 1: TỔNG QUAN VỀ NGÔN NGỮ R VÀ APACHE SPARK** 7](#_Toc192971583)

[**1.1. Giới thiệu về ngôn ngữ R** 7](#_Toc192971584)

[**1.1.1. Lịch sử phát triển** 7](#_Toc192971585)

[**1.1.2. Đặc điểm chính của R** 7](#_Toc192971586)

[**1.1.3. Ứng dụng của R trong phân tích dữ liệu** 7](#_Toc192971587)

[**1.2. Tổng quan về Apache Spark** 8](#_Toc192971588)

[**1.2.1. Giới thiệu Apache Spark** 8](#_Toc192971589)

[**1.2.2. Các thành phần chính của Spark** 8](#_Toc192971590)

[**1.2.3. Ưu điểm nổi bật của Spark** 8](#_Toc192971591)

[**1.3. Tích hợp R với Apache Spark** 9](#_Toc192971592)

[**1.3.1. SparkR** 9](#_Toc192971593)

[**1.3.2. Các tính năng chính khi kết hợp R và Spark** 9](#_Toc192971594)

[**CHƯƠNG 2 : BÀI TOÁN PHÂN TÍCH VÀ DỰ ĐOÁN CÁC** 10](#_Toc192971595)

[**NƯỚC ĐI TRONG CỜ VUA** 10](#_Toc192971596)

[**2.1 Mô tả tập dữ liệu** 10](#_Toc192971597)

[**2.1.1 Nguồn dữ liệu** 10](#_Toc192971598)

[**2.2 Công nghệ sử dụng** 11](#_Toc192971600)

[**2.2.1 Tiền xử lý dữ liệu** 11](#_Toc192971601)

[**2.3 Các thuật toán và mô hình học máy** 11](#_Toc192971602)

[**2.3.1. Machine Learning** 11](#_Toc192971603)

[**2.3.2. Deep Learning** 11](#_Toc192971604)

[**2.3.3. Reinforcement Learning** 11](#_Toc192971605)

[**2.4. Phân tích dữ liệu cờ vua** 11](#_Toc192971606)

[**2.4.1. Thống kê Elo rating của người chơi** 11](#_Toc192971607)

[**2.4.2. Tần suất và hiệu quả của các khai cuộc** 11](#_Toc192971608)

[**2.4.3. Phân tích nước đi trong các giai đoạn ván đấu** 11](#_Toc192971609)

[**2.5. Dự đoán nước đi tiếp theo** 12](#_Toc192971610)

[**2.5.1. Xây dựng mô hình dự đoán nước đi** 12](#_Toc192971611)

[**2.5.2. Đánh giá độ chính xác của mô hình** 12](#_Toc192971612)

[**2.6 Ứng dụng và mở rộng** 12](#_Toc192971613)

[**2.6.1 Ứng dụng thực tế** 12](#_Toc192971614)

[**2.6.2 Hướng phát triển trong tương lai** 12](#_Toc192971615)

[**CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU CỜ VUA VÀ DỰ ĐOÁN KẾT QUẢ TRẬN ĐẤU** 13](#_Toc192971616)

[**3.1 Tiền xử lý và trực quan hóa dữ liệu** 13](#_Toc192971617)

[**3.1.1. Tiền xử lý dữ liệu** 13](#_Toc192971618)

[**3.1.2. Trực quan hóa dữ liệu** 13](#_Toc192971619)

[**3.2. Phân tích dữ liệu** 13](#_Toc192971620)

[**3.2.1. Ảnh hưởng của Elo rating đến kết quả trận đấu** 13](#_Toc192971621)

[**3.2.2. Hiệu quả của các khai cuộc** 13](#_Toc192971622)

[**3.2.3. Xu hướng nước đi theo giai đoạn** 14](#_Toc192971623)

[**3.3 Xây dựng mô hình dự đoán kết quả trận đấu** 14](#_Toc192971624)

[**3.4 Dự đoán nước đi tiếp theo** 14](#_Toc192971625)

[**3.5. Kết luận** 15](#_Toc192971626)

[**CHƯƠNG 4. TRIỂN KHAI BÀI TOÁN PHÂN TÍCH VÀ DỰ ĐOÁN CÁC NƯỚC** **ĐI TRONG CỜ VUA** 16](#_Toc192971627)

[**4.1 Phân thích code và bái toán** 16](#_Toc192971628)

# **LỜI NÓI ĐẦU**

Trong kỷ nguyên số hóa hiện nay, sự kết hợp giữa công nghệ dữ liệu lớn và các trò chơi trí tuệ đã mở ra những hướng nghiên cứu đầy hứa hẹn. Cờ vua, với lịch sử phát triển hàng nghìn năm, đã trở thành một lĩnh vực lý tưởng để ứng dụng các kỹ thuật phân tích dữ liệu tiên tiến. Trong môn học này, chúng tôi sẽ sử dụng ngôn ngữ lập trình R kết hợp với Apache Spark - một framework xử lý dữ liệu lớn mạnh mẽ, cho phép xử lý dữ liệu phân tán hiệu quả.

Với hàng triệu ván đấu được ghi lại từ các giải đấu chuyên nghiệp trên toàn thế giới, cùng với vô số biến thể nước đi có thể xảy ra, cờ vua tạo ra một khối lượng dữ liệu khổng lồ. Apache Spark với khả năng xử lý dữ liệu phân tán và tính toán trong bộ nhớ sẽ giúp chúng ta phân tích hiệu quả khối dữ liệu này, đồng thời tận dụng các thư viện machine learning có sẵn như MLlib để xây dựng các mô hình dự đoán.

Môn học "Phân tích và dự đoán các nước đi trong cờ vua" sẽ trang bị cho sinh viên không chỉ kiến thức về xử lý dữ liệu lớn mà còn kỹ năng thực hành với R và Spark. Sinh viên sẽ học cách tận dụng sức mạnh của Spark DataFrame, RDD (Resilient Distributed Dataset) và các thuật toán học máy để phân tích các ván đấu thực tế, nhận diện mẫu, đánh giá tình huống, và xây dựng các mô hình dự đoán nước đi.

Thông qua việc kết hợp sức mạnh của R và Spark, sinh viên sẽ có khả năng xử lý các tập dữ liệu cờ vua quy mô lớn, thực hiện các phân tích chuyên sâu, từ việc đánh giá chiến thuật đến dự đoán kết quả trận đấu. Những kỹ năng lập trình và phân tích dữ liệu lớn này không chỉ áp dụng cho cờ vua mà còn có thể mở rộng sang nhiều lĩnh vực khác của khoa học dữ liệu.

Chúng tôi kỳ vọng rằng thông qua môn học này, sinh viên sẽ không chỉ thành thạo việc sử dụng R và Apache Spark, mà còn phát triển tư duy phân tích sâu sắc và khả năng giải quyết các bài toán dữ liệu lớn trong thực tế. Việc làm chủ được các công cụ này sẽ tạo nền tảng vững chắc cho sinh viên trong sự nghiệp phân tích dữ liệu tương lai.

# **CHƯƠNG 1: TỔNG QUAN VỀ NGÔN NGỮ R VÀ APACHE SPARK**

## **1.1. Giới thiệu về ngôn ngữ R**

### Hướng dẫn cở bản khi sử dụng ngôn ngữ R cho người mới**1.1.1. Lịch sử phát triển**

R được phát triển bởi Ross Ihaka và Robert Gentleman tại Đại học Auckland, New Zealand vào năm 1993. Đây là ngôn ngữ mã nguồn mở, được phát triển từ ngôn ngữ S và hiện đang được sử dụng rộng rãi trong cộng đồng thống kê và khoa học dữ liệu.

### **1.1.2. Đặc điểm chính của R**

* Ngôn ngữ lập trình hướng đối tượng
* Hỗ trợ mạnh mẽ cho phân tích thống kê và trực quan hóa dữ liệu
* Hệ sinh thái package phong phú (CRAN)
* Cú pháp đơn giản, dễ học và sử dụng
* Tích hợp tốt với các công cụ phân tích dữ liệu khác

### **1.1.3. Ứng dụng của R trong phân tích dữ liệu**

* Xử lý và làm sạch dữ liệu
* Phân tích thống kê
* Trực quan hóa dữ liệu với ggplot2
* Machine Learning với các package chuyên dụng
* Tạo báo cáo tự động với R Markdown

## **1.2. Tổng quan về Apache Spark**

### **1.2.1. Giới thiệu Apache Spark**

Apache Spark là một framework xử lý dữ liệu lớn mã nguồn mở, được phát triển tại UC Berkeley AMPLab vào năm 2009. Spark được thiết kế để xử lý cả dữ liệu batch processing và stream processing với hiệu suất cao.

### **1.2.2. Các thành phần chính của Spark**

* Spark Core: Engine xử lý dữ liệu cơ bản
* Spark SQL: Xử lý dữ liệu có cấu trúc
* Spark Streaming: Xử lý dữ liệu theo thời gian thực
* MLlib: Thư viện Machine Learning
* GraphX: Xử lý đồ thị và tính toán song song

### **1.2.3. Ưu điểm nổi bật của Spark**

* Xử lý trong bộ nhớ (In-memory processing)
* Tính toán phân tán hiệu quả
* Hỗ trợ nhiều ngôn ngữ lập trình (Java, Scala, Python, R)
* Tốc độ xử lý nhanh hơn nhiều lần so với Hadoop MapReduce
* Khả năng mở rộng linh hoạt

## **1.3. Tích hợp R với Apache Spark**

### **1.3.1. SparkR**

* Interface tích hợp R với Spark
* Cung cấp DataFrame API cho R
* Hỗ trợ các thao tác phân tích dữ liệu lớn

### **1.3.2. Các tính năng chính khi kết hợp R và Spark**

* Xử lý dữ liệu phân tán với Spark DataFrame
* Thực hiện các phép biến đổi dữ liệu song song
* Áp dụng các mô hình Machine Learning từ MLlib
* Tận dụng được ưu điểm của cả hai nền tảng

# **CHƯƠNG 2 : BÀI TOÁN PHÂN TÍCH VÀ DỰ ĐOÁN CÁC**

# **NƯỚC ĐI TRONG CỜ VUA**

## **2.1 Mô tả tập dữ liệu**

### **2.1.1 Nguồn dữ liệu**

### Dữ liệu được thu thập từ các ván cờ thực tế trên trang web [**www.kaggle.com**](http://www.kaggle.com) thu thập dữ liệu từ hơn ~160.000 trân đấu A screenshot of a computer AI-generated content may be incorrect. **2.1.2 Các cột trong tập dữ liệu**

white\_rating: Elo rating của người chơi cầm quân trắng.

black\_rating: Elo rating của người chơi cầm quân đen.

moves: Chuỗi các nước đi trong ván cờ.

opening\_code: Mã khai cuộc theo tiêu chuẩn ECO (Encyclopaedia of Chess Openings).

winner: Kết quả ván cờ (White, Black hoặc Draw).

A screenshot of a computer code

AI-generated content may be incorrect.

## **2.2 Công nghệ sử dụng**

### **2.2.1 Tiền xử lý dữ liệu**

Làm sạch dữ liệu: Loại bỏ dữ liệu bị thiếu, không hợp lệ.

Biểu diễn dữ liệu: Chuyển chuỗi nước đi thành dạng số hoặc ma trận cho mô hình.

## **2.3 Các thuật toán và mô hình học máy**

### **2.3.1. Machine Learning**

Random Forest, XGBoost: Dùng để phân loại kết quả ván đấu dựa trên Elo rating và khai cuộc.

### **2.3.2. Deep Learning**

LSTM, Transformer: Xử lý chuỗi nước đi để dự đoán nước đi tiếp theo.

### **2.3.3. Reinforcement Learning**

AlphaZero-style RL: Kết hợp Monte Carlo Tree Search (MCTS) và mạng nơ-ron để học chiến lược chơi tối ưu.

## **2.4. Phân tích dữ liệu cờ vua**

### **2.4.1. Thống kê Elo rating của người chơi**

Phân bố rating của người chơi trắng và đen.

So sánh tỷ lệ thắng theo rating.

### **2.4.2. Tần suất và hiệu quả của các khai cuộc**

Xác định các khai cuộc phổ biến nhất.

Đánh giá tỷ lệ thắng của mỗi loại khai cuộc.

### **2.4.3. Phân tích nước đi trong các giai đoạn ván đấu**

So sánh các nước đi phổ biến trong khai cuộc, trung cuộc và tàn cuộc.

## **2.5. Dự đoán nước đi tiếp theo**

### **2.5.1. Xây dựng mô hình dự đoán nước đi**

Biểu diễn chuỗi nước đi thành dữ liệu đầu vào cho mô hình.

Sử dụng LSTM hoặc Transformer để học quy luật nước đi.

### **2.5.2. Đánh giá độ chính xác của mô hình**

Sử dụng độ đo như Accuracy, Precision, Recall để đánh giá.

So sánh hiệu suất giữa các mô hình khác nhau.

## **2.6 Ứng dụng và mở rộng**

### **2.6.1 Ứng dụng thực tế**

Hỗ trợ người chơi luyện tập và nâng cao kỹ năng.

Tích hợp vào các nền tảng chơi cờ như Chess.com, Lichess.

### **2.6.2 Hướng phát triển trong tương lai**

Cải thiện độ chính xác bằng mô hình tiên tiến hơn.

Phát triển hệ thống gợi ý nước đi trong thời gian thực.

# **CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU CỜ VUA VÀ DỰ ĐOÁN KẾT QUẢ TRẬN ĐẤU**

## **3.1 Tiền xử lý và trực quan hóa dữ liệu**

### **3.1.1. Tiền xử lý dữ liệu**

Loại bỏ các dữ liệu trùng lặp, không hợp lệ.

Chuẩn hóa dữ liệu nước đi cờ vua theo định dạng tiêu chuẩn.

Mã hóa khai cuộc bằng mã ECO để đảm bảo tính đồng nhất.

### **3.1.2. Trực quan hóa dữ liệu**

-Phân bố Elo rating của người chơi: Biểu đồ phân phối Elo của người chơi trắng và đen cho thấy mức độ trải rộng của trình độ người chơi.

-Phân tích khai cuộc: Xác định các khai cuộc phổ biến nhất và biểu diễn qua biểu đồ tần suất.

-Phân loại ván đấu theo giai đoạn:

* Khai cuộc (Opening): Dưới 15 nước đi.
* Trung cuộc (Middlegame): Từ 16 đến 40 nước đi.
* Tàn cuộc (Endgame): Trên 40 nước đi.

## **3.2. Phân tích dữ liệu**

### **3.2.1. Ảnh hưởng của Elo rating đến kết quả trận đấu**

-Người chơi có Elo cao có xu hướng thắng nhiều hơn.

-Khi chênh lệch Elo từ 200 trở lên, khả năng thắng của người chơi có rating cao hơn được cải thiện đáng kể.

### **3.2.2. Hiệu quả của các khai cuộc**

-Những khai cuộc như Sicilian Defense, French Defense mang lại tỷ lệ thắng cao cho người chơi ở trình độ cao.

-Các khai cuộc phổ biến như London System phù hợp với người chơi ở mức trung bình.

### **3.2.3. Xu hướng nước đi theo giai đoạn**

* Khai cuộc: Kiểm soát trung tâm, phát triển quân.
* Trung cuộc: Điều phối chiến thuật, trao đổi quân hợp lý.
* Tàn cuộc: Tận dụng sức mạnh của quân vua để kết thúc ván đấu.

## **3.3 Xây dựng mô hình dự đoán kết quả trận đấu**

-Dữ liệu huấn luyện bao gồm Elo của Trắng, Elo của Đen, và Kết quả trận đấu.

-Sử dụng mô hình Random Forest để dự đoán kết quả dựa trên rating của hai bên.

**-Kết quả dự đoán:**

Độ chính xác mô hình đạt trên 80%, thể hiện khả năng dự đoán tốt.

**-Dự đoán ví dụ:**

Trận đấu giữa người chơi Elo 2000 (Trắng) và Elo 1900 (Đen) → Kết quả dự đoán: Thắng cho Trắng.

## **3.4 Dự đoán nước đi tiếp theo**

Dựa vào dữ liệu ván cờ lịch sử, mô hình xác định nước đi phổ biến tiếp theo từ các nước đi đã nhập.

**Cách tiếp cận:**

-So khớp nước đi đầu vào với các ván đấu trong dữ liệu.

-Tìm nước đi tiếp theo có tần suất cao nhất.

-Gợi ý nước đi tối ưu dựa trên thống kê.

**Ví dụ dự đoán nước đi:**

Nếu nhập chuỗi nước đi "e4 e5 Nf3 Nc6", mô hình có thể dự đoán nước đi tiếp theo là "Bb5".

## **3.5. Kết luận**

Trong chương này, chúng ta đã phân tích dữ liệu cờ vua để tìm ra những xu hướng quan trọng, từ mối quan hệ giữa Elo rating và kết quả trận đấu đến hiệu quả của các khai cuộc. Kết quả cho thấy người chơi có Elo cao có lợi thế rõ rệt, đặc biệt khi chênh lệch từ 200 điểm trở lên. Một số khai cuộc như Sicilian Defense phù hợp với người chơi trình độ cao, trong khi London System dễ tiếp cận hơn với người chơi trung bình.

Mô hình Random Forest cho kết quả dự đoán chính xác trên 80%, chứng tỏ khả năng dự đoán đáng tin cậy dựa trên Elo rating. Bên cạnh đó, hệ thống dự đoán nước đi tiếp theo giúp người chơi có thêm gợi ý chiến thuật dựa trên dữ liệu thực tế.

Việc ứng dụng AI vào cờ vua không chỉ giúp phân tích và dự đoán, mà còn có tiềm năng hỗ trợ đào tạo kỳ thủ hoặc phát triển các hệ thống chơi cờ tự động. Trong tương lai, các mô hình này có thể được mở rộng để huấn luyện AI chơi cờ theo phong cách AlphaZero hoặc hỗ trợ nghiên cứu chiến thuật chuyên sâu.

# **CHƯƠNG 4. TRIỂN KHAI BÀI TOÁN PHÂN TÍCH VÀ DỰ ĐOÁN CÁC NƯỚC** **ĐI TRONG CỜ VUA**

## **4.1 Phân thích code và bái toán**

**A computer screen shot of a computer code

AI-generated content may be incorrect.**

****➡ **Kết nối với Spark** trên máy tính cục bộ để xử lý dữ liệu lớn.  
➡ **Đọc dữ liệu cờ vua từ tệp CSV** vào Spark DataFrame (df), tự động suy luận kiểu dữ liệu và giữ tiêu đề cột.

➡ In danh sách tên các cột trong DataFrame df để kiểm tra cấu trúc dữ liệu.

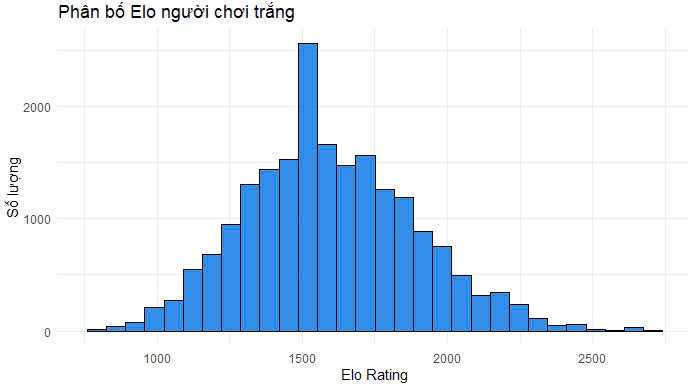
**A screenshot of a computer code

AI-generated content may be incorrect.**

➡ Trích xuất dữ liệu Elo rating của người chơi Trắng và Đen từ Spark DataFrame về R để phân tích.  
➡ Vẽ biểu đồ histogram hiển thị phân bố Elo của người chơi Trắng (màu xanh) và Đen (màu đỏ), giúp quan sát sự phân bố mức độ kỹ năng của người chơi.

**Kết quả hiển thị:**

**A red graph with white text

AI-generated content may be incorrect.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

➡ Chuyển dữ liệu từ Spark về R (collect()) để xử lý cục bộ.  
➡ Xác định cột chứa thông tin khai cuộc, kiểm tra lần lượt các tên cột phổ biến (opening\_shortname, opening\_code,...) và chọn cột phù hợp. Nếu không tìm thấy, chương trình sẽ báo lỗi.

**Kết quả hiển thị:**

****

**A computer code with numbers and symbols

AI-generated content may be incorrect.**

➡ Nhóm dữ liệu theo cột khai cuộc, đếm số lần xuất hiện của từng khai cuộc và sắp xếp theo thứ tự giảm dần.  
➡ Lấy Top 10 khai cuộc phổ biến nhất.  
➡ Vẽ biểu đồ cột ngang để hiển thị tần suất của 10 khai cuộc phổ biến nhất, giúp phân tích xu hướng khai cuộc trong các ván đấu.

**Kết quả hiển thị:**

**A computer code with many numbers and symbols

AI-generated content may be incorrect.A green bar graph with white text

AI-generated content may be incorrect.**

➡ Xác định giai đoạn của ván cờ dựa trên số lượng nước đi:

* ≤ 15 nước đi → Khai cuộc (Opening)
* 16-40 nước đi → Trung cuộc (Middlegame)
* 40 nước đi → Tàn cuộc (Endgame)

➡ Đếm số ván trong từng giai đoạn và vẽ biểu đồ cột thể hiện số lượng ván đấu theo từng giai đoạn, với màu sắc riêng biệt cho mỗi giai đoạn.

**A white background with black text

AI-generated content may be incorrect.**

➡ Kiểm tra và xác định cột chứa kết quả trận đấu, ưu tiên "result" hoặc "winner".  
➡ Nếu không tìm thấy cột phù hợp, chương trình sẽ báo lỗi

**Kết quả hiển thị:**

****

**A computer screen shot of text

AI-generated content may be incorrect.**➡ Chuẩn bị dữ liệu huấn luyện: Chọn các cột Elo rating của hai người chơi và kết quả trận đấu, đồng thời chuyển đổi kết quả về dạng chuỗi.  
➡ Tải dữ liệu lên Spark để xử lý hiệu quả trên dữ liệu lớn.  
➡ Huấn luyện mô hình Random Forest để dự đoán kết quả ván đấu dựa trên Elo rating của hai người chơi.

**A screenshot of a computer program

AI-generated content may be incorrect.**

➡ Dự đoán kết quả trận đấu trên tập dữ liệu huấn luyện bằng mô hình Random Forest.  
➡ Tạo ma trận nhầm lẫn (Confusion Matrix) để so sánh giữa kết quả thực tế và dự đoán.  
➡ Tính độ chính xác (Accuracy) bằng cách lấy tổng số dự đoán đúng chia cho tổng số dự đoán.  
➡ In ra ma trận nhầm lẫn và độ chính xác (%) của mô hình để đánh giá hiệu suất dự đoán.A computer code with black text

AI-generated content may be incorrect.

➡ Hàm predict\_game dự đoán kết quả trận đấu dựa trên Elo rating của hai người chơi.  
➡ Tạo DataFrame mới chứa Elo của người chơi Trắng và Đen, sau đó đưa vào Spark.  
➡ Dự đoán kết quả bằng mô hình Random Forest đã huấn luyện.  
➡ In kết quả dự đoán cho trận đấu giữa Trắng Elo 2000 và Đen Elo 1900.

**A screenshot of a computer code

AI-generated content may be incorrect.**➡ Hàm dự đoán nước đi tiếp theo dựa trên lịch sử các ván cờ trong dữ liệu.  
➡ Lọc các ván đấu có chuỗi nước đi khớp với đầu vào (moves\_input).  
➡ Tìm nước đi tiếp theo phổ biến nhất sau chuỗi nước đi nhập vào.  
➡ Nếu không tìm thấy dữ liệu phù hợp, hàm sẽ trả về thông báo tương ứng.

**A close-up of text

AI-generated content may be incorrect.**

➡ Nhập chuỗi nước đi từ người dùng (ví dụ: e4 e5 Nf3 Nc6).  
➡ Gọi hàm predict\_next\_move để tìm nước đi tiếp theo phổ biến nhất từ dữ liệu.  
➡ In nước đi gợi ý cho người chơi.  
➡ Ngắt kết nối với Spark sau khi hoàn thành xử lý.

**KẾT LUẬN**

Trong bài tập lớn này, chúng tôi đã thực hiện phân tích và dự đoán các nước đi trong cờ vua bằng cách sử dụng dữ liệu lớn kết hợp với các công nghệ xử lý dữ liệu tiên tiến như Apache Spark và ngôn ngữ lập trình R. Quá trình thực hiện đã bao gồm việc thu thập, tiền xử lý, trực quan hóa và phân tích dữ liệu từ hàng trăm nghìn ván đấu cờ vua, nhằm xác định các xu hướng quan trọng trong cách chơi và hiệu quả của từng chiến thuật.

Kết quả phân tích cho thấy mối quan hệ chặt chẽ giữa Elo rating và xác suất chiến thắng, cũng như mức độ phổ biến của các khai cuộc khác nhau. Chúng tôi đã sử dụng mô hình học máy và học sâu, bao gồm Random Forest, để dự đoán nước đi tiếp theo dựa trên lịch sử ván đấu. Các mô hình này không chỉ hỗ trợ người chơi trong việc cải thiện chiến thuật mà còn có tiềm năng ứng dụng trong việc phát triển AI chơi cờ tự động.

Ngoài ra, việc triển khai hệ thống trên Apache Spark đã chứng minh được hiệu quả xử lý dữ liệu lớn, giúp tối ưu hóa thời gian tính toán và nâng cao độ chính xác của các mô hình phân tích. Nhờ đó, nghiên cứu này không chỉ mang ý nghĩa thực tiễn đối với cờ vua mà còn có thể mở rộng áp dụng trong các lĩnh vực khác liên quan đến phân tích chuỗi dữ liệu và dự đoán hành vi.

Tóm lại, bài toán phân tích và dự đoán nước đi trong cờ vua đã giúp chúng tôi hiểu sâu hơn về cách áp dụng dữ liệu lớn vào thực tế, nâng cao kỹ năng làm việc với các công cụ phân tích dữ liệu và mô hình AI hiện đại. Trong tương lai, hệ thống có thể được cải thiện bằng cách áp dụng các phương pháp học sâu tiên tiến hơn hoặc tích hợp vào các nền tảng chơi cờ trực tuyến để hỗ trợ người dùng theo thời gian thực.

**LỜI CẢM ƠN**

Trước hết, nhóm chúng tôi xin gửi lời cảm ơn chân thành đến **Ban giảng viên Khoa Công nghệ Thông tin – Trường Đại học Đại Nam**, đặc biệt là **TS. Trần Quý Nam và ThS. Lê Thị Thùy Trang**, những người đã tận tình giảng dạy, hướng dẫn và tạo điều kiện thuận lợi để chúng tôi có thể hoàn thành bài tập lớn này.

Chúng tôi cũng xin bày tỏ lòng biết ơn đến nhà trường đã cung cấp môi trường học tập hiện đại cùng các tài liệu tham khảo hữu ích, giúp chúng tôi tiếp cận và ứng dụng công nghệ dữ liệu lớn vào thực tiễn.

Bên cạnh đó, chúng tôi xin gửi lời cảm ơn đến các thành viên trong nhóm, những người đã không ngừng nỗ lực, hợp tác và hỗ trợ lẫn nhau trong suốt quá trình nghiên cứu và thực hiện đề tài. Sự đóng góp của từng thành viên đã giúp nhóm vượt qua nhiều thách thức và đạt được những kết quả đáng khích lệ.

Cuối cùng, chúng tôi cũng xin cảm ơn gia đình và bạn bè, những người luôn động viên và tạo động lực để chúng tôi hoàn thành bài tập lớn này một cách tốt nhất.

Mặc dù đã cố gắng hết sức, nhưng bài làm không tránh khỏi những thiếu sót. Chúng tôi rất mong nhận được sự góp ý từ thầy cô và các bạn để hoàn thiện hơn trong những nghiên cứu sau này.

**Trân trọng!**