**TRƯỜNG ĐẠI HỌC VINH**

**VIỆN KỸ THUẬT VÀ CÔNG NGHỆ**

****

**BÁO CÁO**

**ÁP DỤNG THUẬT TOÁN KNN**

**ĐỂ PHÁT HIỆN GIAN LẬN THẺ TÍN DỤNG**

|  |  |
| --- | --- |
|  |  |
| Sinh viên: | Lê Văn Mạnh, 215748020110398 |
|  |  |
|  |  |

**Nghệ An,** **03/2025**

MỤC LỤC

[1. Giới thiệu 3](#_Toc192261681)

[2. Dữ liệu và phương pháp 3](#_Toc192261682)

[***2.1. Mô tả dữ liệu*** 3](#_Toc192261683)

[***2.2. Tiền xử lý dữ liệu*** 3](#_Toc192261684)

[3. Xây dựng mô hình KNN 4](#_Toc192261685)

[***3.2. Huấn luyện mô hình*** 4](#_Toc192261686)

[***3.3. Điều chỉnh tham số*** 4](#_Toc192261687)

[4. Kết luận và hướng phát triển 5](#_Toc192261688)

[5. Tài liệu tham khảo 5](#_Toc192261689)

**1. Giới thiệu**

Gian lận trong giao dịch tài chính là một vấn đề quan trọng trong lĩnh vực an ninh tài chính. Với sự phát triển của công nghệ, các hình thức gian lận ngày càng trở nên tinh vi hơn, gây tổn thất lớn cho các tổ chức tài chính. Do đó, việc xây dựng các mô hình phát hiện gian lận trở thành một nhiệm vụ cấp thiết.

Mục tiêu của báo cáo này là xây dựng một mô hình học máy để phân loại các giao dịch có dấu hiệu gian lận dựa trên tập dữ liệu thực tế. Chúng tôi sử dụng thuật toán KNN để thực hiện phân loại, đồng thời đánh giá hiệu suất của mô hình qua các chỉ số quan trọng như Accuracy, Precision, Recall và F1-score.

**2. Dữ liệu và phương pháp**

***2.1. Mô tả dữ liệu***

Tập dữ liệu sử dụng trong bài toán là creditcard\_2023.csv, chứa thông tin về các giao dịch tài chính. Tập dữ liệu có tổng cộng 30 đặc trưng và một nhãn (Class), trong đó:

**V1 - V28**: Các đặc trưng ẩn được tạo bởi phương pháp PCA.

**Amount:** Số tiền của giao dịch.

**Class:** Nhãn của giao dịch, với giá trị 0 là hợp pháp và 1 là gian lận.

***2.2. Tiền xử lý dữ liệu***

Đoạn code sau kiểm tra số lượng giá trị bị thiếu trong mỗi cột của dữ liệu:

Ảnh có chứa ảnh chụp màn hình, văn bản, Phông chữ, Đồ họa

Nội dung do AI tạo ra có thể không chính xác.

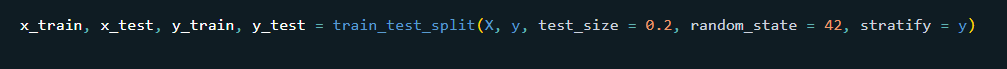
Nếu có dữ liệu bị thiếu, có thể xử lý bằng cách loại bỏ hoặc điền giá trị thay thế.

Chuẩn hóa dữ liệu: Dữ liệu được chuẩn hóa bằng StandardScaler để đưa tất cả các đặc trưng về cùng một khoảng giá trị, giúp thuật toán hoạt động hiệu quả hơn:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Dữ liệu được chia thành 80% cho huấn luyện và 20% cho kiểm tra:



**3. Xây dựng mô hình KNN**

### 3.1. Giới thiệu thuật toán KNN

Thuật toán K-Nearest Neighbors (KNN) là một trong những thuật toán đơn giản và hiệu quả trong học máy. Nguyên tắc hoạt động của KNN dựa trên khoảng cách giữa các điểm dữ liệu:

* Một điểm dữ liệu mới được phân loại dựa trên số lượng điểm lân cận gần nhất.
* Sử dụng khoảng cách Euclidean để đo độ tương đồng.
* Giá trị của K (số điểm lân cân được xem xét) ảnh hướng lớn đến kết quả phân loại.

Công thức khoảng cách Euclidean giữa hai điểm dữ liệu x và y trong không gian n chiều:

d(x, y) =

***3.2. Huấn luyện mô hình***

Xây dựng và huấn luyện mô hình với K = 5:

Ảnh có chứa ảnh chụp màn hình, văn bản, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

KNeighborsClassifier tìm kiếm 5 điểm dữ liệu gần nhất để xác định nhãn của giao dịch mới.

***3.3. Điều chỉnh tham số***

Tìm giá trị K tối ưu bằng cách thử nghiệm nhiều giá trị khác nhau và tính toán tỷ lệ lỗi:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Biểu đồ Accuracy Score sẽ giúp chọn giá trị K tốt nhất.

**4. Kết luận và hướng phát triển**

Mô hình KNN đã được áp dụng để phát hiện gian lận tài chính. Tuy nhiên, thuật toán KNN có một số nhược điểm như:

* Nhạy cảm với dữ liệu không cân bằng.
* Hiệu suất giảm khi kích thước dữ liệu lớn.

**5. Tài liệu tham khảo**

[1] Scikit-learn Documentation: https://scikit-learn.org/

[2] Fraud Detection Dataset: https://www.kaggle.com/datasets