TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VIỆN TRÍ TUỆ NHÂN TẠO

-----*-----**

BÁO CÁO MÔN HỌC KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

ĐỀ TÀI

PHÂN LOẠI SỬ DỤNG CÔNG CỤ HỖ TRỢ KNIME ĐỂ XÁC ĐỊNH NHỮNG KHÁCH HÀNG TIỀM NĂNG TRONG CATCH THE PINK FLAMINGO

Nhóm sinh viên thực hiện:

1. Nguyễn Nam Dương
2. Chu Hữu Đăng Trường
3. Nguyễn Tiến Dũng

Giảng viên hướng dẫn: TS. Trần Hồng Việt

ThS. Ngô Minh Hương

HÀ NỘI, 5/2024

MỞ ĐẦU

Công nghệ Big data đã đạt đến đỉnh cao trong việc thực hiện các chức năng của nó. Trong tháng 8/2015 Big data đã vượt ra khỏi bảng xếp hạng những công nghệ mới nổi Cycle Hype của Gartner và tạo ra một tiếng vang lớn cho xu hướng công nghệ của thế giới. Big data chứa trong mình rất nhiều thông tin quý giá mà nếu mà trích xuất thành công, nó sẽ giúp rất nhiều trong nhiều lĩnh vực như y tế, giao thông, giáo dục, …

Chính vì thế, những framework giúp việc xử lý dữ liệu lớn như Hadoop, Spark, và KNIME trở nên vô cùng quan trọng. Những công cụ này không chỉ giúp các tổ chức thu thập, lưu trữ, và xử lý lượng dữ liệu khổng lồ mà còn tối ưu hóa việc phân tích để khai thác những giá trị tiềm ẩn từ dữ liệu đó.

Trong bài tập lớn này, chúng em đã chọn KNIME để nghiên cứu về hành vi người dùng trên 1 trò chơi làm báo kết thúc môn học của mình. Tên đề tài là: “Phân loại trong KNIME để xác định những khách hàng tiềm năng trong Catch the Pink Flamingo”

Báo cáo gồm 4 chương:

Chương 1: Tổng quan về KNIME

Chương 2: Tổng quan về dữ liệu Catch the Pink - Flamingo

Chương 3: Dự đoán kiểu người mua hàng sử dụng KNIME

Chương 4: Phân tích kết quả mô hình, kết luận và hướng phát triển

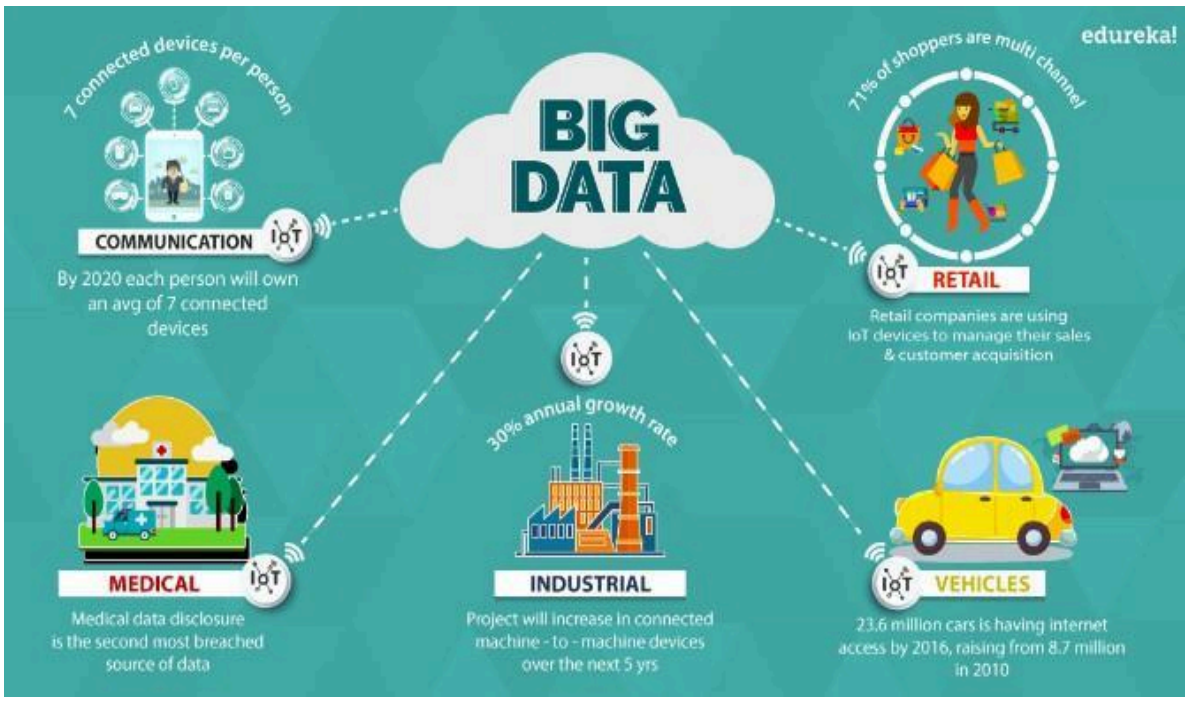
CHƯƠNG 1: TỔNG QUAN VỀ KNIME

1.1 Tổng quan về dữ liệu lớn

Định nghĩa dữ liệu lớn:

- *Theo wikipedia*: Dữ liệu lớn là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này.
- *Theo Gartner*: Dữ liệu lớn là những nguồn thông tin có đặc điểm chung khối lượng lớn, tốc độ nhanh và dữ liệu định dạng dưới nhiều hình thức khác nhau,do đó muốn khai thác được phải đòi hỏi phải có hình thức mới để đưa ra quyết định khám phá và tối ưu hóa quy trình.

Dữ liệu đến từ rất nhiều nguồn khác nhau:

- 

Hình 1. Minh họa nguồn gốc của dữ liệu.

Một số lợi ích có thể mang lại như: Cắt giảm chi phí, tiết kiệm thời gian và giúp tối ưu hóa sản phẩm, hỗ trợ con người đưa ra những quyết định đúng và hợp lý hơn.

Một số đặc trưng của dữ liệu lớn:

(1) *Khối lượng lớn (Volume)*: Khối lượng dữ liệu rất lớn và đang ngày càng tăng lên, tính đến 2014 thì có thể trong khoảng vài trăm terabyte.

(2) *Tốc độ (Velocity)*: Khối lượng dữ liệu gia tăng rất nhanh.

(3) *Đa dạng (Variety):* Ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc (tài liệu, blog,

hình ảnh,...)

(4) *Độ tin cậy/chính xác(Veracity):* Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của bigdata.

(5) *Giá trị(Value):* Giá trị thông tin mang lại.

1.2 Tổng quan về KNIME

KNIME (Konstanz Information Miner) là một nền tảng mã nguồn mở được phát triển bởi công ty KNIME GmbH tại Đức. Nền tảng này được thiết kế để hỗ trợ phân tích dữ liệu và khai phá dữ liệu, nổi bật với giao diện người dùng trực quan và khả năng mở rộng mạnh mẽ. KNIME cho phép người dùng xây dựng và triển khai các quy trình phân tích dữ liệu phức tạp mà không cần kỹ năng lập trình sâu, nhờ vào việc kéo thả các nút (nodes) trên giao diện đồ họa. Các nút này bao gồm các chức năng từ đơn giản như đọc dữ liệu, xử lý và chuyển đổi dữ liệu, cho đến các thuật toán học máy phức tạp và trực quan hóa dữ liệu.

Dưới đây là một số lợi ích của KNIME:

- **Mã nguồn mở và miễn phí**: KNIME được phát triển dựa trên mã nguồn mở, cho phép người dùng tùy chỉnh và mở rộng chức năng theo nhu cầu cụ thể mà không phải trả phí bản quyền.
- **Giao diện kéo thả**: Người dùng có thể xây dựng quy trình phân tích dữ liệu một cách trực quan bằng cách kéo thả các nút trên giao diện, giảm thiểu nhu cầu về kiến thức lập trình.
- **Khả năng mở rộng**: KNIME hỗ trợ tích hợp với nhiều ngôn ngữ lập trình và công cụ khác như Python, R, SQL, và nhiều thư viện học máy khác, giúp tăng cường khả năng phân tích và xử lý dữ liệu.
- **Hỗ trợ đa dạng dữ liệu**: KNIME có thể xử lý nhiều định dạng dữ liệu khác nhau, từ các tệp văn bản, cơ sở dữ liệu SQL, NoSQL, đến dữ liệu từ các nguồn dịch vụ web.
- **Thư viện phong phú**: Nền tảng này cung cấp một thư viện khổng lồ với hàng ngàn nút phục vụ cho các mục đích khác nhau, từ xử lý dữ liệu, phân tích thống kê, học máy, đến trực quan hóa dữ liệu.
- **Khả năng tích hợp tốt**: KNIME có thể dễ dàng tích hợp với các hệ thống doanh nghiệp và các công cụ phân tích khác, tạo điều kiện thuận lợi cho việc chia sẻ và sử dụng dữ liệu trong toàn tổ chức.
- **Cộng đồng hỗ trợ lớn**: Với một cộng đồng người dùng và các nhà phát triển rộng lớn, người dùng KNIME có thể dễ dàng tìm kiếm sự trợ giúp, chia sẻ kinh nghiệm và cập nhật những cải tiến mới nhất.

Ngoài những lợi ích chung trên, KNIME có lợi ích riêng với việc phân tích dữ liệu lớn như sau:

- **Xử lý dữ liệu lớn**: KNIME được thiết kế để xử lý các tập dữ liệu lớn một cách hiệu quả, với khả năng quản lý và phân tích dữ liệu trên các cụm máy tính và tích hợp với Hadoop và Spark.
- **Hiệu suất cao**: Nền tảng này tối ưu hóa các quy trình xử lý và phân tích dữ liệu, giúp giảm thiểu thời gian xử lý và cải thiện hiệu suất công việc.

- **Khả năng tự động hóa**: KNIME cho phép tự động hóa các quy trình phân tích dữ liệu, giúp giảm thiểu công việc thủ công và tăng cường độ chính xác của kết quả phân tích.
- **Phân tích dữ liệu nâng cao**: Với sự hỗ trợ của các thuật toán học máy và khai phá dữ liệu tiên tiến, KNIME giúp khám phá các mô hình, xu hướng và thông tin ẩn trong các tập dữ liệu lớn.
- **Trực quan hóa dữ liệu**: KNIME cung cấp nhiều công cụ trực quan hóa dữ liệu, giúp người dùng dễ dàng hiểu và trình bày kết quả phân tích dữ liệu lớn một cách sinh động và rõ ràng.
- **Bảo mật và quản lý dữ liệu**: Nền tảng này đảm bảo an toàn và bảo mật cho dữ liệu, đồng thời cung cấp các công cụ quản lý dữ liệu hiệu quả, phù hợp với các tiêu chuẩn bảo mật của ngành.

Nhờ những tính năng và lợi ích trên, KNIME không chỉ là một công cụ hữu ích cho các nhà phân tích dữ liệu mà còn là một giải pháp mạnh mẽ cho việc xử lý và phân tích dữ liệu lớn, giúp các tổ chức tận dụng tối đa giá trị từ dữ liệu của mình.

CHƯƠNG 2: TỔNG QUAN VỀ DỮ LIỆU CATCH THE PINK - FLAMINGO

Dự đoán người dùng nào có khả năng mua các mặt hàng có giá trị lớn khi chơi Catch the Pink Flamingo là kiến thức quý giá cần có đối với Eglence vì mua hàng trong ứng dụng là nguồn doanh thu chính. Trong nhiệm vụ này, bạn sẽ phân tích dữ liệu có sẵn để phân loại người dùng là người mua các mặt hàng có giá trị lớn ("HighRollers") so với người mua các mặt hàng rẻ tiền ("PennyPinchers"). Các mặt hàng có giá trị lớn là những mặt hàng có giá trên 5 USD và các mặt hàng rẻ tiền là những mặt hàng có giá từ 5 USD trở xuống.

2.1 Phân tích dữ liệu gốc

Chúng em đã được cung cấp tệp dữ liệu "combined_data.csv" đã được tạo cho bài tập này. Các trường dữ liệu của file csv như sau:

- userId: Id người dùng
- userSessionId: ID phiên sử dụng của người dùng
- team_level: Cấp độ đội của người dùng
- platformType: Nền tảng thiết bị của người dùng
- count_gameclicks: Tổng số lần click trong phiên sử dụng
- count_hits: Tổng số lần click trúng
- count_buyid: Số lần mua item trong phiên bản
- avg_price: Giá tính trung bình của vật phẩm mua trong phiên

	userId	userSessionId	teamLevel	platformType	count_gameclicks	count_hits	count_buyId	avg_price
0	812	5648	1	android	69	8	NaN	NaN
1	1658	5649	1	iphone	31	5	NaN	NaN
2	1589	5650	1	iphone	26	2	NaN	NaN
3	1863	5651	1	android	35	4	NaN	NaN
4	937	5652	1	android	39	0	1.0	1.0

Hình 2. Minh họa dữ liệu ban đầu

Một số phân tích của chúng em về dữ liệu ban đầu:

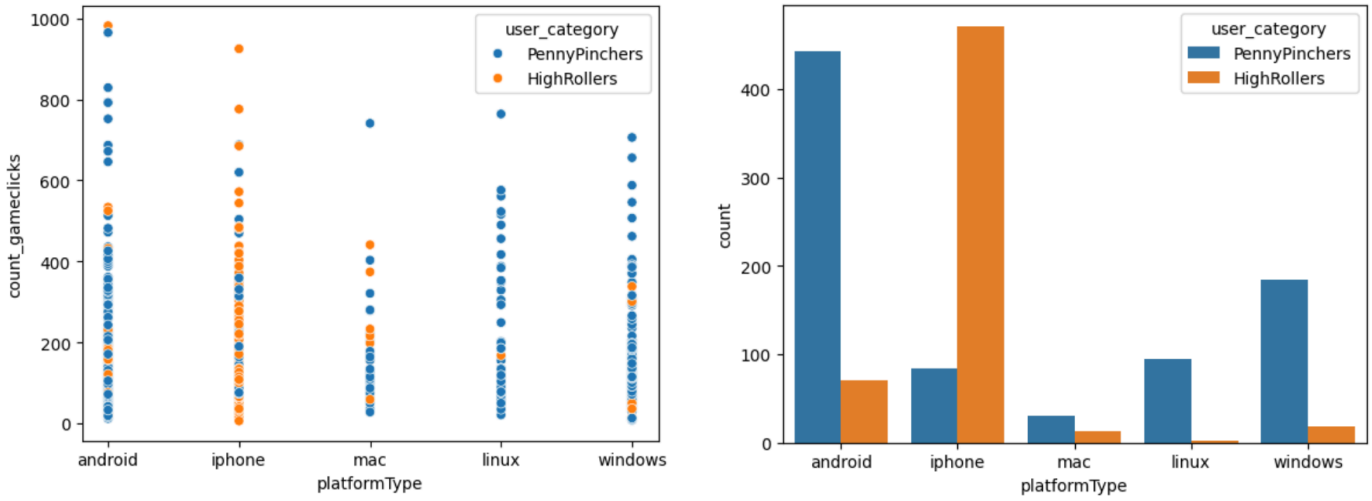
- Tổng doanh thu trong game từ việc mua trang bị: 21406
- Số lượng trang bị có thể mua: 6
- Số phiên tham gia chơi: 4619
- Số phiên người dùng mua trang bị: 1411

Từ yêu cầu đề bài, chúng em lọc dữ liệu và chỉ sử dụng những dữ liệu có avg_price và count_buyid không NULL. Những sample nào có avg_price > 5 sẽ được đánh nhãn spender_type là "HighRoller", còn những user còn lại là "PennyPincher". Cuối cùng, loại bỏ những cột không sử dụng trong việc phân loại người dùng. Bộ dữ liệu sau cùng gồm 1411 hàng, gồm các cột: ["platformType", "count_gameclicks", "count_buyid", "spender_type"]

	platformType	count_gameclicks	count_hits	spender_type
0	iphone	129	9	HighRoller
1	android	90	10	PennyPincher
2	iphone	51	8	HighRoller
3	android	47	5	PennyPincher
4	android	46	7	PennyPincher

Hình 3. Minh họa dữ liệu sau khi đã gộp lại

Trực quan hóa dựa trên các cột này, chúng em ngay lập tức thấy yếu tố đặc biệt chiếm phần lớn doanh thu đó là nền tảng người dùng sử dụng. Số lượng người dùng trên nền tảng iPhone phần lớn là HighRoller.



Hình 4. Một số hình ảnh trực quan hóa dữ liệu

Bảng số liệu tổng kết từ dữ liệu ban đầu:

platformType	teamLevel	count_gameclicks	count_hits	count_buyId	total_price
android	4.769981	144.284600	16.062378	1.654971	6.676452
iphone	4.915315	142.816216	17.401802	1.684685	21.975135
linux	4.354167	157.604167	17.510417	1.666667	2.187292
mac	5.045455	165.090909	18.363636	1.613636	7.931818
windows	4.955665	144.093596	16.054187	1.773399	3.985123

Hình 5. Bảng tổng kết dữ liệu ban đầu

Từ bảng tổng kết này chúng em thấy rằng các nền tảng có chỉ số trung bình count_gameclicks và count_hits tương đương nhau nên 2 yếu tố này rõ ràng không có tác động nhiều đến việc chi tiền mua trang bị của người dùng. Bên cạnh đó như đã nói iPhone vẫn là yếu tố có ảnh hưởng lớn nhất đến việc mua vật phẩm khi nhìn vào cột total_price (trung bình tổng số tiền đã được chi từ người dùng trong các phiên).

2.2 Khảo sát kết quả phân loại với một số mô hình Học Máy

Ngoài ra, chúng em đã thử việc phân loại người trên một số mô hình Học Máy như Decision Tree, Naive Bayes, KNN, SVC trước khi sử dụng KNIME và thấy kết quả như sau:

```
Prediction for Decision Tree :  0.8017699115044248
Accuracy for Naive Bayes :  0.7893805309734513
Accuracy for K neighbors :  0.5362831858407080
Accuracy for SVM :  0.5716814159292035
```

Hình 6. Kết quả một số mô hình Học Máy

Có thể thấy Decision Tree cho kết quả phân loại tốt nhất với 80% chính xác, từ kết quả này kết hợp với yêu cầu đề bài, chúng em đã thống nhất sử dụng công cụ KNIME với workflow là mô hình Decision Tree cũng như tập trung vào trường platformType để có thể đạt hiệu quả phân loại tốt nhất.

CHƯƠNG 3: DỰ ĐOÁN KIỂU NGƯỜI MUA HÀNG SỬ DỤNG KNIME

3.1 Ý tưởng

➢ **Nhiệm vụ:** Dự đoán một người dùng là PennyPincher (người mua tiết kiệm) hay HighRoller (người mua hàng tiềm năng).
➢ Ý tưởng:

* Từ dữ liệu những người dùng đã mua hàng, chia tập train/test theo tỉ lệ 60/40.

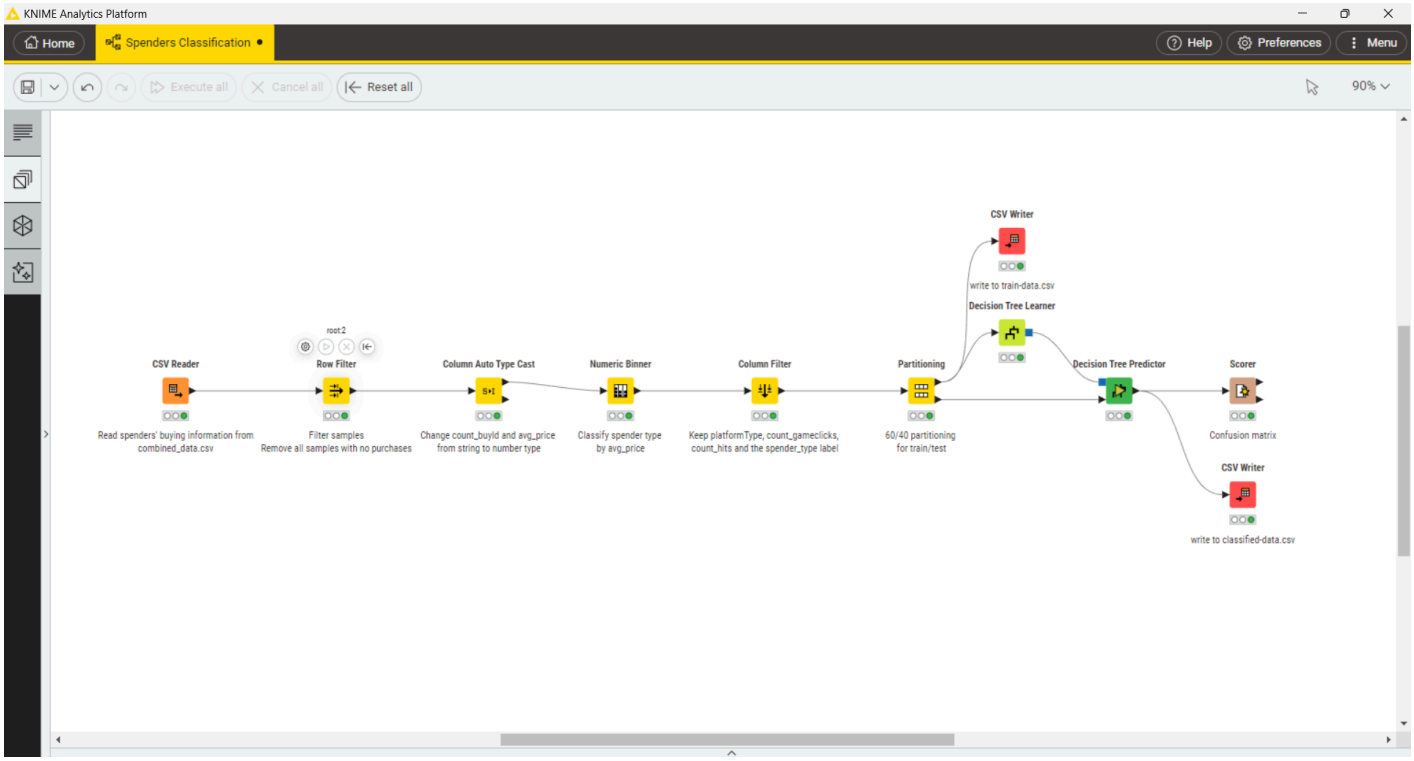
* Phân tích dữ liệu + sử dụng decision tree để đưa ra mô hình dự đoán loại người dùng.

3.2 Demo dự đoán dùng KNIME

➢ ***Dữ liệu đầu vào****:* Là danh sách các hàng lưu dưới dạng file .csv. Mỗi hàng là dữ liệu của từng người chơi như đã đề cập ở trên, cụ thể bao gồm:

- userId: Id người dùng
- userSessionId: ID phiên sử dụng của người dùng
- team_level: Cấp độ đội của người dùng
- platformType: Nền tảng thiết bị của người dùng
- count_gameclicks: Tổng số lần click trong phiên sử dụng
- count_hits: Tổng số lần click trúng
- count_buyid: Số lần mua item trong phiên bản
- avg_price: Giá tính trung bình của vật phẩm mua trong phiên

➢ ***Triển khai:***

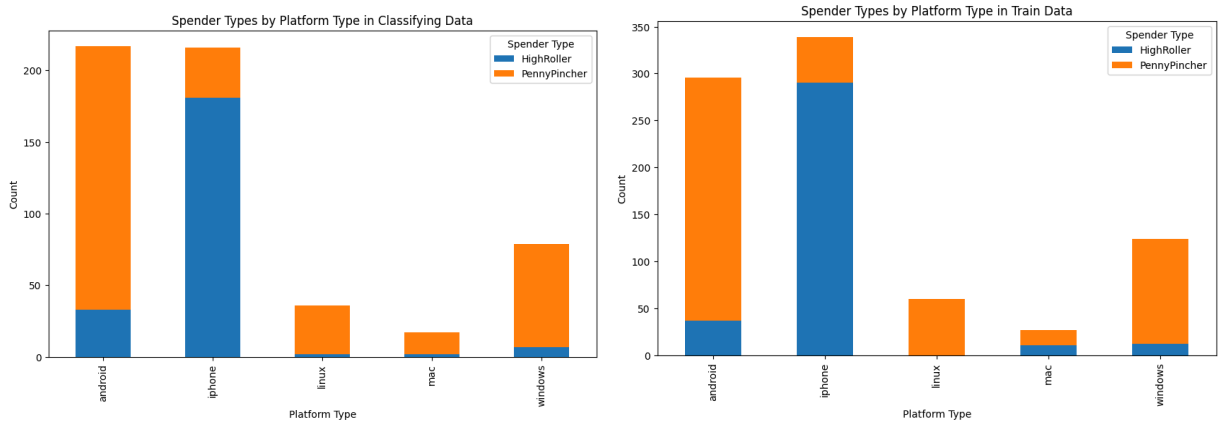


1. Đọc dữ liệu từ file combined_data.csv (file dữ liệu đầu vào) bằng node CSV Reader.
2. Lọc các sample không thực hiện thao tác mua sản phẩm, đổi kiểu dữ liệu của avg_price và count_buyId về dạng Numerical (node Row Filter và Colum Auto Type Cast).
3. Tạo label cho dữ liệu dựa theo tiêu chí phân loại người dùng với HighRoller là người mua các sản phẩm có giá lớn hơn 5 USD, còn lại là PennyPincher (node Numeric Binner).
4. Lọc bỏ những cột không cần thiết cho classification. Chỉ giữ lại platformType, count_gameclicks, count_hits và spender_type (label).
5. Chia dữ liệu theo tỷ lệ train/test là 60/40 (node Partitioning đưa 2 output, train data và test data).
6. Train decision tree với input là train data đã chia ở node trước đó (node Decision Tree Learner).
7. Thực hiện dự đoán trên tập test với model đã train ở node Learner trước đó (node Decision Tree Predictor với 2 input là decision tree đã được học và test data đã chia ở node Partitioning)
8. Dự đoán kết quả với output của node Predictor (node Scorer).
9. Node Scorer có 2 output là confusion matrix và accuracy statistic của input đưa v

CHƯƠNG 4: PHÂN TÍCH KẾT QUẢ MÔ HÌNH, KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

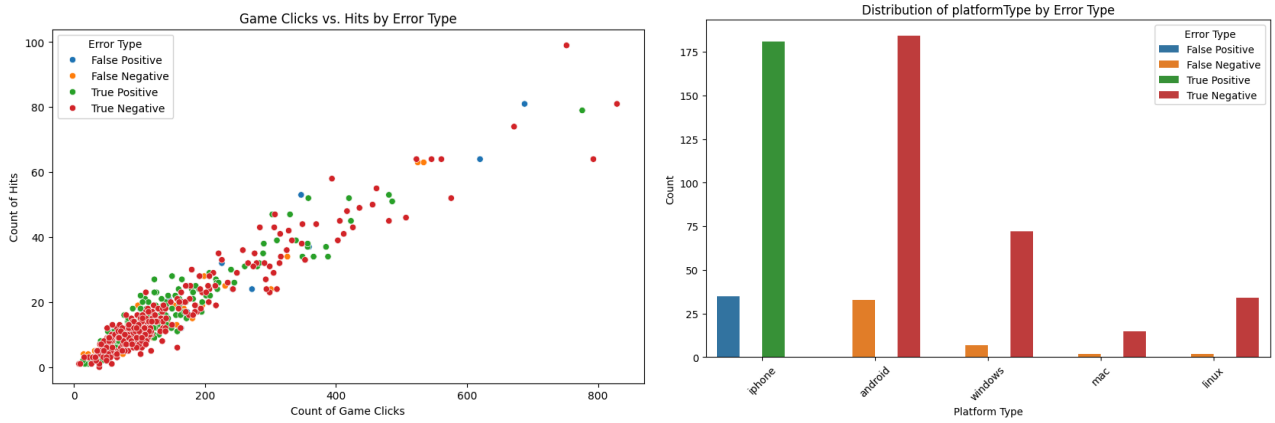
4.1 Phân tích mô hình

- Kết quả mô hình: cả tập train và tập test đều cho thấy các người dùng iPhone sẽ là nhóm khách hàng tiềm năng, còn count_gameclicks và count_hits gần như không ảnh hưởng.



Hình 7. Một số kết quả của mô hình

- Ngoài ra, vì thuật toán Decision Tree có xu hướng cắt nhánh mạnh, nên chúng em dự đoán mô hinh sẽ dự đoán nhầm những người dùng iPhone PennyPincher sang HighRoller và những người chơi mà không sử dụng nền tảng iPhone mà là HighRoller sang PennyPincher.



-

Hình 8. Một số kết quả sau khi phân tích các dự đoán sai

True Positives: Dự đoán High Roller, thực tế High Roller
True Negatives: Dự đoán Penny Pincher, thực tế Penny Pincher

False Positives: Dự đoán High Roller, thực tế Penny Pincher
False Negatives: Dự đoán Penny Pincher, thực tế High Roller

4.2 Kết luận.

KNIME nổi bật với tính hiệu quả so với Spark và Hadoop nhờ vào giao diện người dùng thân thiện và khả năng xử lý dữ liệu mà không cần kỹ năng lập trình chuyên sâu. Trong khi Spark và Hadoop yêu cầu người dùng phải thành thạo các ngôn ngữ lập trình như Scala, Java, hoặc Python, KNIME cho phép xây dựng và triển khai quy trình phân tích dữ liệu thông qua giao diện kéo thả trực quan. Điều này giúp giảm thiểu thời gian học tập và triển khai, làm cho KNIME trở thành lựa chọn lý tưởng cho những người không chuyên về lập trình nhưng vẫn muốn thực hiện các phân tích phức tạp. KNIME tích hợp sẵn các nút chức năng mạnh mẽ, từ xử lý dữ liệu, học máy đến trực quan hóa, giúp tăng hiệu quả công việc và tối ưu hóa quy trình phân tích dữ liệu.

Nhóm em đã đạt được một số kết quả như sau:

➢ Hiểu tổng quan về KNIME

➢ Triển khai bài toán phân loại trên KNiME

➢ Xây dựng thành công chương trình demo cho đề tài.

➢ Đánh giá chương trình.

Tuy nhiên kết quả vẫn còn một số hạn chế:

➢ Việc phân tích dữ liệu vẫn còn ở mức nhỏ.

➢ Chương trình demo mới chỉ áp dụng đúng cho dữ liệu đó, dữ liệu khác chưa hỗ trợ được.

4.3 Hướng phát triển.

➢ Áp dụng kiến thức về Big data, apache hadoop, cải tiến và xây dựng ứng dụng phân tích dữ liệu lớn hơn và vào nhiều lĩnh vực khác.

➢ Trong quá trình hoàn thành bài tập lớn, nhóm em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của thầy và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn.

PHỤ LỤC: NHIỆM VỤ CỦA CÁC THÀNH VIÊN

Thành viên	Nội dung thực hiện
Nguyễn Nam Dương	- Thực hiện workflow KNIME và demo - Trình bày bài tập lớn - Hỗ trợ nội dung report
Chu Hữu Đăng Trường	- Phân tích dữ liệu và kết quả - Viết report
Nguyễn Tiến Dũng	- Phân tích dữ liệu gốc - Làm slide