



Trường Đại học Công nghệ, ĐHQGHN

Hà Nội, ngày 28 tháng 5 năm 2024

Classifying in KNIME to identify big spenders in Catch the Pink Flamingo



Nhóm 5

22022644@vnu.edu.vn

Nguyễn Tiến Dũng

22022512@vnu.edu.vn

Nguyễn Nam Dương

22022505@vnu.edu.vn

Chu Hữu Đăng Trường

MỤC LỤC

1. Bài toán
đặt ra

2. Công cụ
KNIME

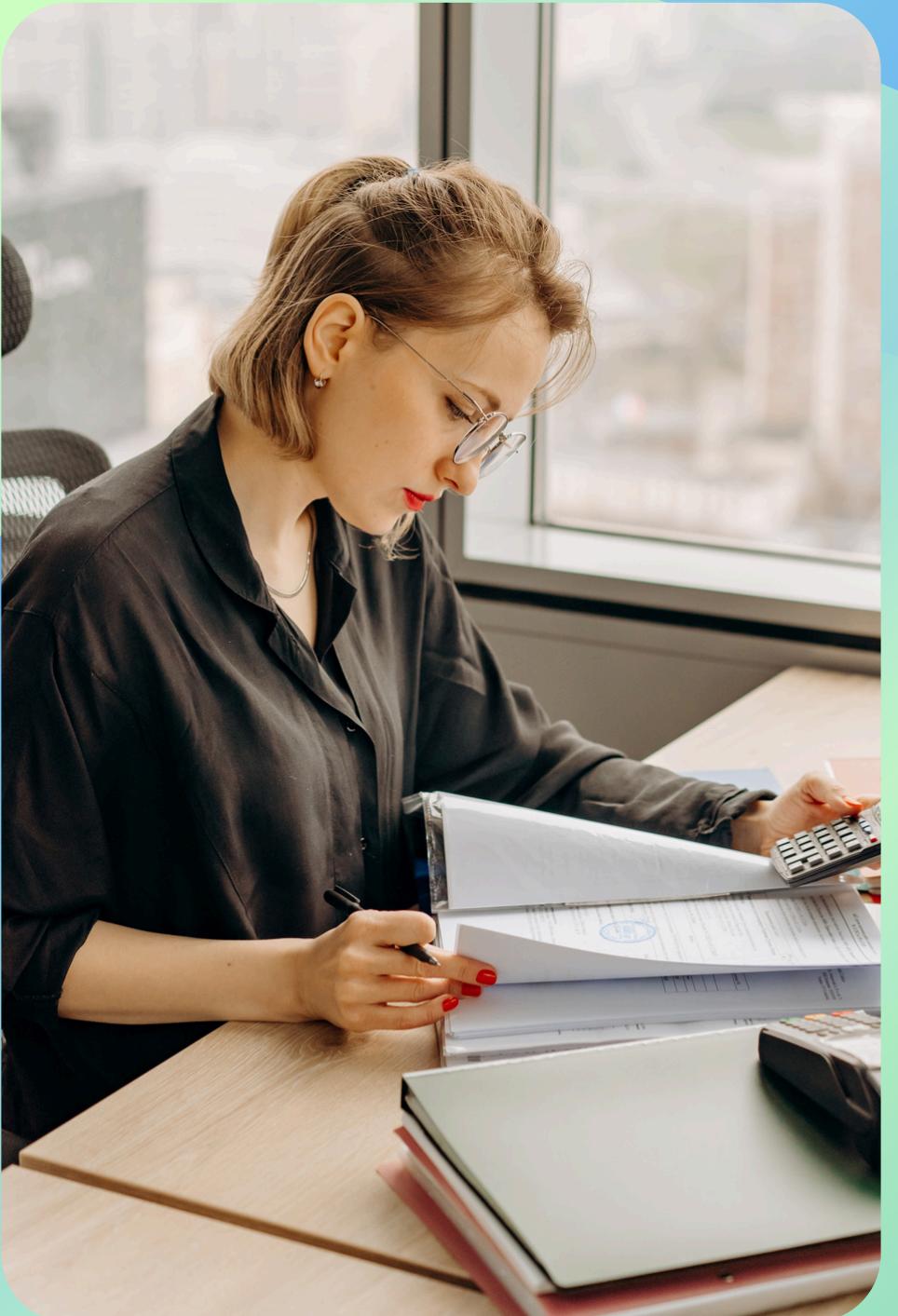
3. Tổng quan
dữ liệu

4. Phân tích
dữ liệu

5. Workflow

6. Phân loại
với KNIME

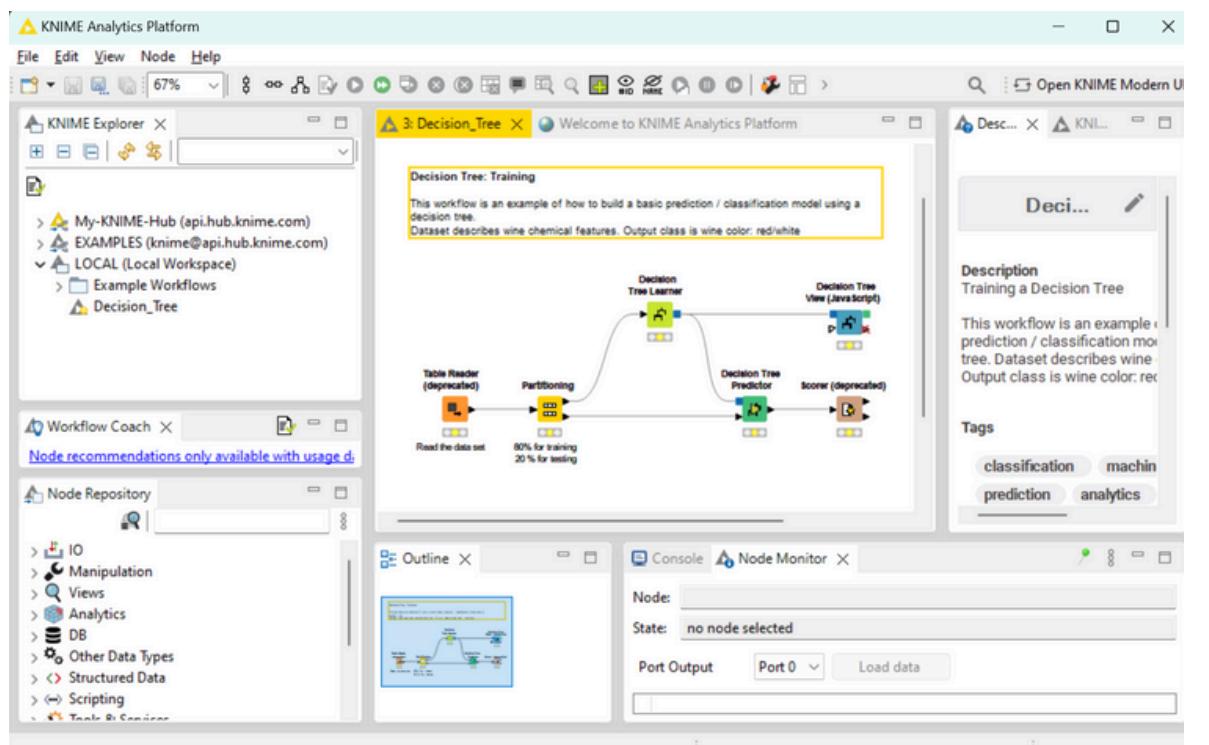
7. Đề xuất
cải tiến



BÀI TOÁN ĐẶT RA

- Đầu vào: Bộ dữ liệu người chơi trong tựa game “Catch the Pink Flamingo”.
- Yêu cầu: Phân tích, phân loại dữ liệu và có sử dụng công cụ KNIME.
- Đầu ra: Các đề xuất giúp tăng doanh thu cho nhà sản xuất.

CÔNG CỤ KNIME



Là một nền tảng mã nguồn mở dành cho phân tích dữ liệu và học máy.



Hỗ trợ các quy trình xử lý dữ liệu từ thu thập đến trực quan hóa.



Giao diện kéo thả trực quan và dễ sử dụng (workflows).



Tích hợp hàng trăm node và plugin mở rộng.



Hỗ trợ cả phiên bản mã nguồn mở và thương mại.

TỔNG QUAN DỮ LIỆU

Dataset:
`catchThePinkFlamingo.csv`

Cột	Nội dung
userID	ID người chơi
userSessionID	ID các phiên người chơi
team_Level	Cấp độ của team
platformType	Nền tảng người chơi sử dụng
count_gameclicks	Số lần click trong phiên của người chơi
count_hits	Số lần click trúng
count_buyID	Số lần mua item trong phiên
avg_price	Giá mua trung bình trong phiên

PHÂN TÍCH DỮ LIỆU

Tổng doanh thu từ items trong game:

21407

Số lượng items có thể mua trong game:

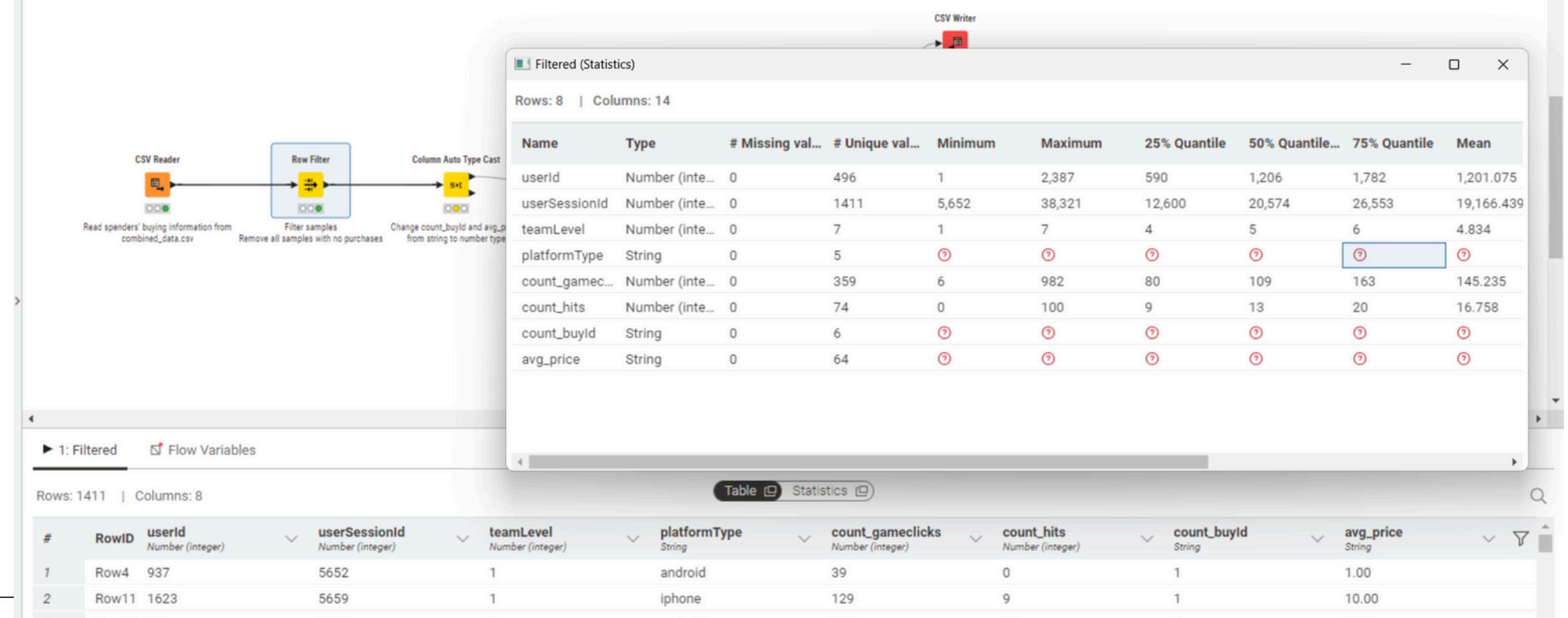
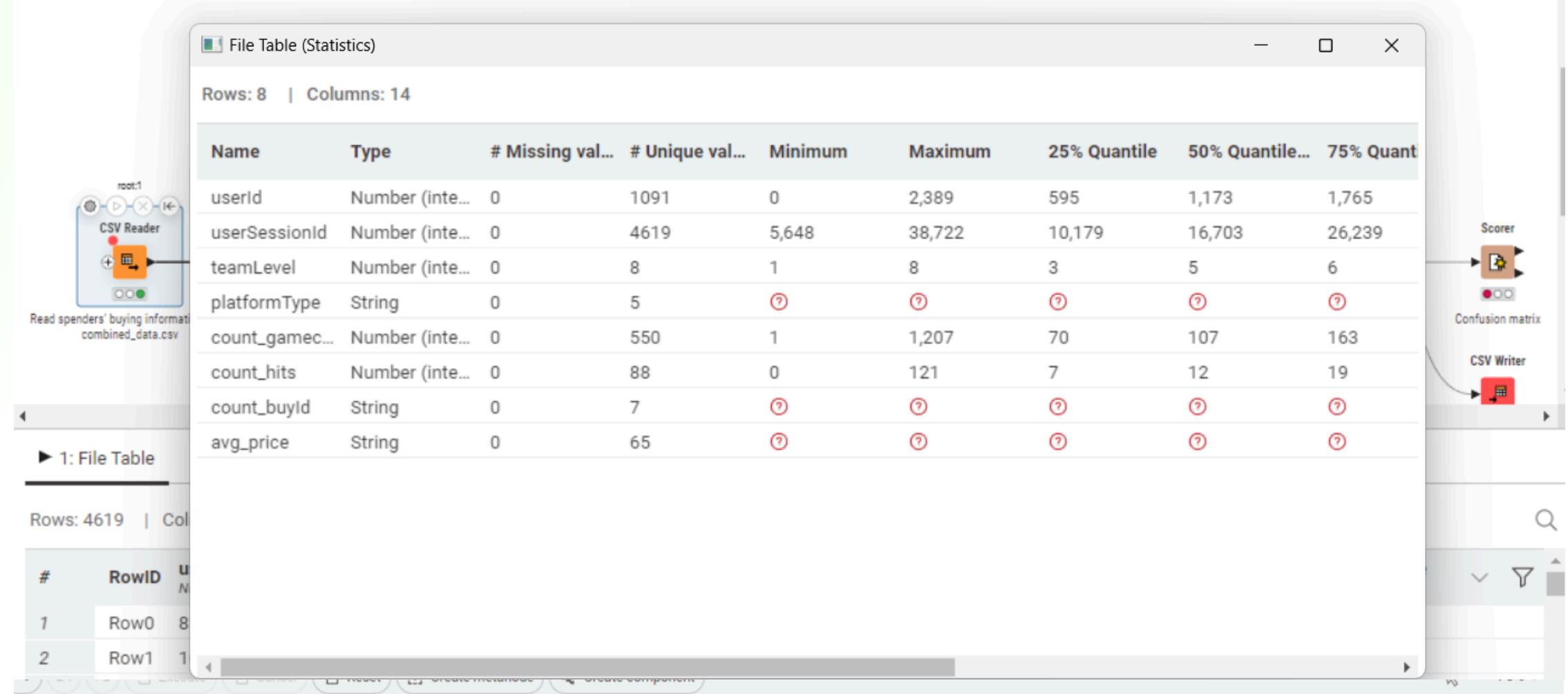
6

Tổng số lượt chơi:

4619

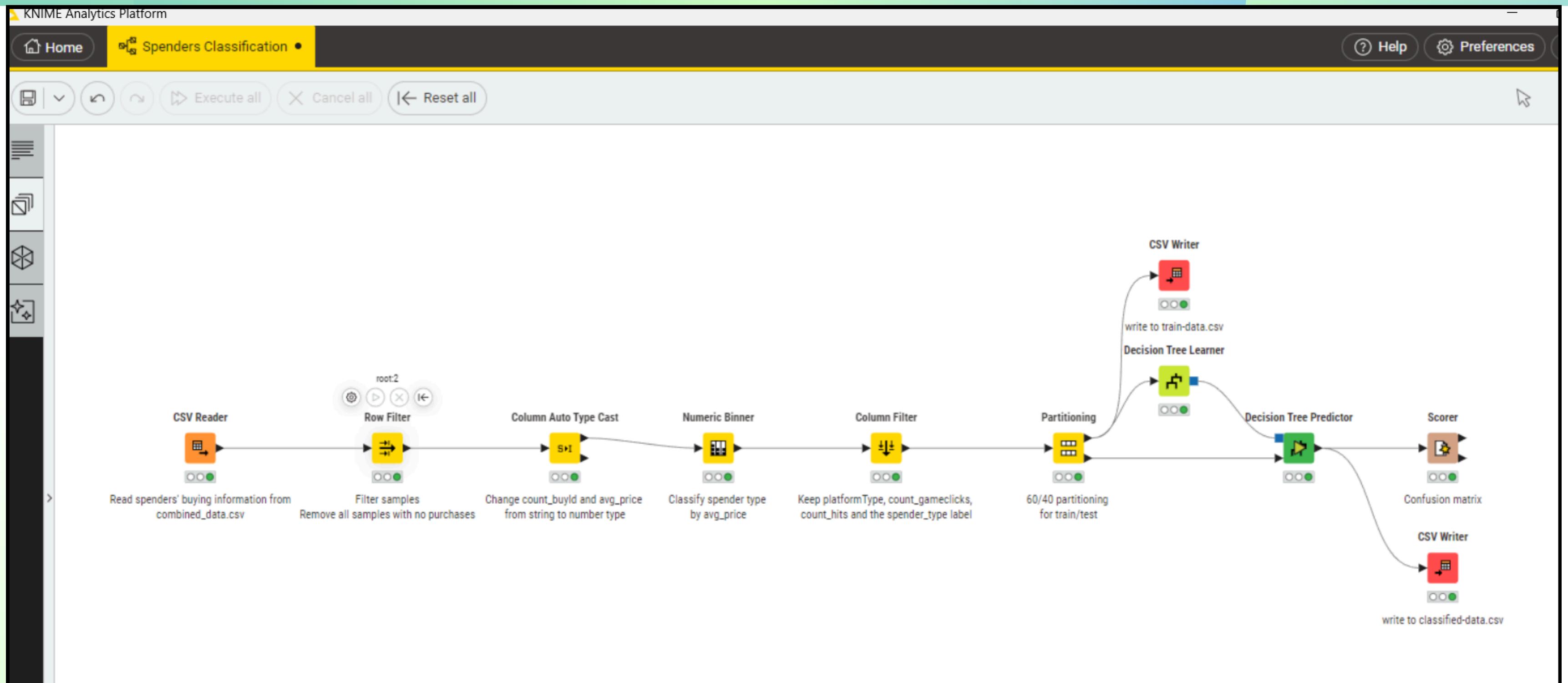
Số lượt chơi có mua item:

1411



WORKFLOW

Big Data Capstone Project



PHÂN LOẠI VỚI KNIME

Big Data Capstone Project

The screenshot displays the KNIME Analytics Platform interface. The title bar shows "KNIME Analytics Platform" and the project name "Spenders Classification". The main window is titled "Decision Tree Model - 3:8 - Decision Tree Learner".

The left sidebar contains icons for Home, Recent Projects, Help, Preferences, and Menu. The main workspace shows a workflow diagram:

- A "Decision Tree Learner" node is connected to a "CSV Writer" node.
- The "CSV Writer" node is connected to a file output labeled "write to train-data.csv".
- A "partitioning" node is connected to the "Decision Tree Learner" node.
- A "partitioning train/test" node is also present in the workspace.

The right side of the screen shows the "PMMI: TreeModel" view, which displays the PMML code for the decision tree model. The PMML code is as follows:

```
PMMI version="4.2" xmlns="http://www.dmg.org/PMML-4_2"
<?xml version="1.0" encoding="UTF-8"?>
<PMML>
    <Header copyright="user" />
    <DataDictionary numberFields="4" />
    <TreeModel modelName="DecisionTree" functionName="classification" splitCharacteristic="multiSplit" missingValueStrategy="lastPrediction" noTrueChildStrategy="returnNullPrediction">
        <MiningSchema />
        <Node id="0" score="PennyPincher" recordCount="846.0">
            <True>
                <ScoreDistribution value="HighRoller" recordCount="350.0" />
                <ScoreDistribution value="PennyPincher" recordCount="496.0" />
            <Node id="1" score="HighRoller" recordCount="339.0">
                <SimplePredicate field="platformType" operator="equal" value="iphone" />
                <ScoreDistribution value="HighRoller" recordCount="290.0" />
                <ScoreDistribution value="PennyPincher" recordCount="49.0" />
            <Node id="70" score="PennyPincher" recordCount="296.0">
                <SimplePredicate field="platformType" operator="equal" value="android" />
                <ScoreDistribution value="HighRoller" recordCount="37.0" />
                <ScoreDistribution value="PennyPincher" recordCount="259.0" />
            <Node id="133" score="PennyPincher" recordCount="27.0">
                <SimplePredicate field="platformType" operator="equal" value="mac" />
                <ScoreDistribution value="HighRoller" recordCount="11.0" />
                <ScoreDistribution value="PennyPincher" recordCount="16.0" />
            <Node id="140" score="PennyPincher" recordCount="124.0">
                <SimplePredicate field="platformType" operator="equal" value="windows" />
                <ScoreDistribution value="HighRoller" recordCount="12.0" />
                <ScoreDistribution value="PennyPincher" recordCount="112.0" />
            <Node id="165" score="PennyPincher" recordCount="60.0">
                <SimplePredicate field="platformType" operator="equal" value="linux" />
                <ScoreDistribution value="HighRoller" recordCount="0.0" />
                <ScoreDistribution value="PennyPincher" recordCount="60.0" />
            </Node>
        </Node>
    </TreeModel>
</PMML>
```

PHÂN LOẠI VỚI KNIME

Big Data Capstone Project

KNIME Analytics Platform

Spenders Classification •

Home Help Preferences Menu

Execute Cancel Reset Create metanode

Decision Tree View - 3:9 - Decision Tree Predictor

PennyPincher (496/846)

Table:

Category	%	n
HighRoller	41.4	350
PennyPincher	58.6	496
Total	100.0	846

platformType = iphone platformType = android platformType = mac platformType = windows platformType = linux

HighRoller (290/339) PennyPincher (259/296) PennyPincher (16/27) PennyPincher (112/124) PennyPincher (60/60)

Table:

Category	%	n
HighRoller	85.5	290
PennyPincher	14.5	49
Total	40.1	339

Category	%	n
HighRoller	12.5	37
PennyPincher	87.5	259
Total	35.0	296

Category	%	n
HighRoller	40.7	11
PennyPincher	59.3	16
Total	3.2	27

Category	%	n
HighRoller	9.7	12
PennyPincher	90.3	112
Total	14.7	124

Category	%	n
HighRoller	0.0	0
PennyPincher	100.0	60
Total	7.1	60

Column Filter

CSV Writer write to train-data.csv

Decision Tree Learner

Partitioning 60/40 partitioning for train/test

Decision Tree Predictor

Scorer

Confusion matrix

CSV Writer write to classified-data.csv

Add comment

1: Classified Data Flow Variables

Rows: 565 | Columns: 5

Table Statistics

#	RowID	platformType	count_gameclicks	count_hits	spender_type	Prediction (spender_type)
21	Row...	windows	41	5	PennyPincher	PennyPincher
22	Row...	linux	34	3	PennyPincher	PennyPincher
23	Row...	android	66	12	PennyPincher	PennyPincher
24	Row...	iphone	272	24	PennyPincher	HighRoller
25	Row...	windows	54	4	PennyPincher	PennyPincher
26	Row...	windows	8	1	PennyPincher	PennyPincher
27	Row...	iphone	22	1	HighRoller	HighRoller

PHÂN LOẠI VỚI KNIME

Big Data Capstone Project

KNIME Analytics Platform

Spenders Classification •

Home Help Preferences Menu

Execute Cancel Reset Create metanode Create component 76%

Accuracy statistics (Table)

Rows: 3 | Columns: 11

#	RowID	TruePositive Number (integer)	FalsePositive Number (integer)	TrueNegative Number (integer)	FalseNegative Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Penn...	305	44	181	35	0.897	0.874	0.897	0.804	0.885	0.86	0.706
2	High...	181	35	305	44	0.804	0.838	0.804	0.897	0.821	0.86	0.706
3	Over...	?	?	?	?	?	?	?	?	?	0.86	0.706

CSV Reader → Row Filter → Column Auto Type Cast → Numeric Binner → Column Filter → Partitioning → Decision Tree Predictor → Scorer → Confusion matrix

1: Confusion matrix 2: Accuracy statistics Flow Variables

Table Statistics

Rows: 2 | Columns: 2

#	RowID	PennyPincher Number (integer)	HighRoller Number (integer)
1	Penn...	305	35
2	High...	44	181

KẾT LUẬN

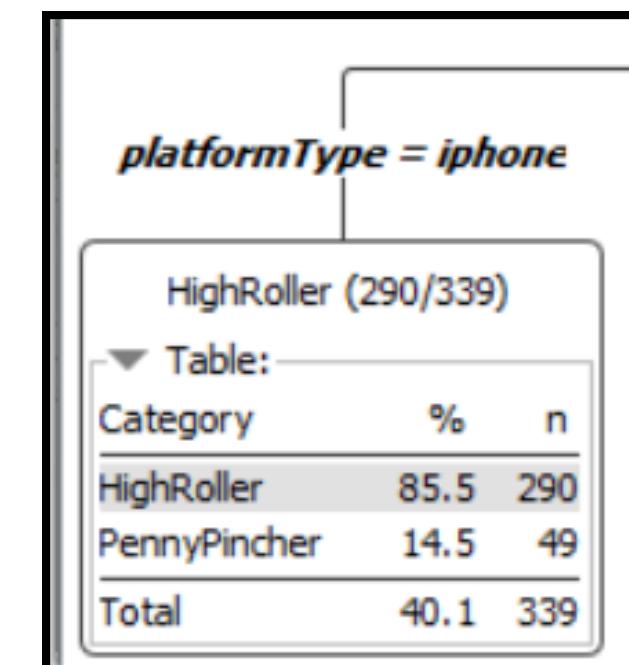
Big Data Capstone Project

Sau khi sử dụng KNIME chạy mô hình Decision Tree với tập dữ liệu trên ta thấy:

- Mô hình phân loại đạt độ chính xác **86%**.
- Lỗi phân loại chủ yếu đến từ việc người dùng iPhone được phân loại là HighRoller dù là PennyPincher và ngược lại.

#	RowID	Accuracy Number (dou...)
1	Penn...	?
2	High...	?
3	Over...	0.86

- Nền tảng người dùng sử dụng có ảnh hưởng lớn nhất đến kết quả phân loại khi đa số HighRollers là người dùng **iPhone**, các nền tảng khác đa số là PennyPinchers.



ĐỀ XUẤT CẢI TIẾN

1. Tập trung quảng cáo các sản phẩm, nhất là các sản phẩm cao cấp, cho các người dùng iPhone.
2. Với những phiên bản tiếp theo của trò chơi, tập trung vào việc tối ưu hóa với người dùng iOS, vì đây là nhóm khách hàng mang lại lợi nhuận cao nhất.



Trường Đại học Công nghệ, ĐHQGHN



Bộ môn: Kỹ thuật và công
nghệ phân tích Dữ liệu lớn

**Thank you for your time!
Reach out to us for questions.**