คู่มือการวิเคราะห์ข้อมูลด้วย

Rapid Miner Studio 9

ในเล่มนี้จะแนะนำให้รู้จักการวิเคราะห์ข้อมูลด้วยวิธี

Data mining ซึ่ง สามารถทำได้ง่าย ๆ จาก Software RapidMiner Studio9

โดยนางสาวนฤชล โรจนบุรานนท์

@Rodjanaburanon

rapidminer

คู่มือการใช้งาน

Rapid Miner Studio 9

หัวข้อต่าง ๆ

1. แนะนำ **Data Mining** และ **RapidMiner Studio**.....**1**
2. เริ่มต้นใช้งาน **RapidMiner Studio** 9.....**4**
3. องค์ประกอบของ **RapidMiner Studio** 9.....**6**
4. ตัวอย่างการสร้างโมเดล **Decision Tree**.....**10**
5.การจัดการข้อมูล **Data Manipulation**.....**19**
6.การทดสอบการทำนาย.....**36**
7. เอกสารอ้างอิง.....

การทำเหมืองข้อมูล (Data Mining)

เนื่องด้วยปัจจุบันเป็นยุคที่ข้อมูลสารและสนเทศมีความสำคัญ การเผยแพร่และสื่อสารข้อมูลข่าวสาร ที่ตรงกับความต้องการของผู้ใช้จึงเป็นสิ่งจำเป็น การประยุกต์เทคโนโลยีสารสนเทศเพื่อช่วยในการสื่อสารข้อมูลจำนวนมากให้แก่ผู้ ใช้ เช่น การให้บริการเว็บไซต์เพื่อเผยแพร่ข้อมูลข่าวสารและแลกเปลี่ยนความรู้ จึงเป็นเครื่องมือที่สำคัญในการสื่อสารข้อมูลถึงผู้ใช้จำนวนมาก ดังนั้นการศึกษาเกี่ยวกับพฤติกรรมของผู้ใช้บริการเว็บไซต์ จะช่วยให้องค์กรสามารถนำข้อมูลมาใช้ในการวางแผนพัฒนาเว็บไซต์ ให้ตรงกับความต้องการใช้งานหรือใช้ในการวางแผนกลยุทธ์ เพื่อสร้างความได้เปรียบทางการแข่งขัน

การทำเหมืองข้อมูล (Data Mining)

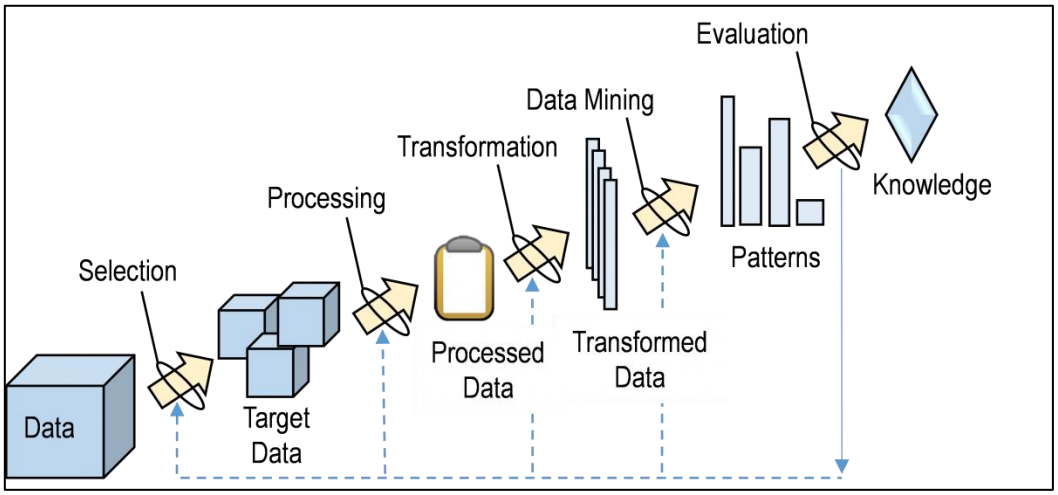
การทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมาย ข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

วิวัฒนาการของการทำเหมืองข้อมูล

- ปี 1960 Data Collection คือ การนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือและป้องกันการสูญ หายได้เป็นอย่างดี
- ปี 1980 Data Access คือ การนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ต่อกันในข้อมูลเพื่อประโยชน์ใน การนำไปวิเคราะห์ และการตัดสินใจอย่างมีคุณภาพ
- ปี 1990 Data Warehouse & Decision Support คือ การรวบรวมข้อมูลมาจัดเก็บลงไปในฐานข้อมูลขนาดใหญ่โดยครอบคลุมทุกด้านของ องค์กร เพื่อช่วยสนับสนุนการตัดสินใจ
- ปี 2000 Data Mining คือ การนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล โดยการสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

ขั้นตอนการทำเหมืองข้อมูล



ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ ประกอบด้วยขั้นตอนดังนี้

- Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป
- Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน
- Data Selection เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้
- Data Transformation เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน
- Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่
- Pattern Evaluation เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล
- Knowledge Representation เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ

RapidMiner Studio 9

ซอฟต์แวร์ RapidMiner Studio 7 แรกเริ่มพัฒนาขึ้นจากบริษัทที่ชื่อว่า Rapid-I ในประเทศเยอรมนีและเมื่อช่วงปลายปี 2013 ที่ผ่านมาได้รับทุนก้อนโตจากนักลงทุนในประเทศสหรัฐอเมริกาจึงเปลี่ยนชื่อบริษัทจาก Rapid-I เป็น RapidMiner แทน และย้ายสำนักงานใหญ่มาอยู่ประเทศสหรัฐอเมริกา เราสามารถดาวน์โหลดซอฟต์แวร์ RapidMiner Studio 9 ซึ่งเป็นเวอร์ชันปัจจุบันได้จากเว็บไซต์ https://rapidminer.com

ข้อดีของซอฟต์แวร์ RapidMiner Studio 9 สรุปได้ดังนี้

- รองรับการใช้งานไฟล์ได้หลายประเภท เช่น ไฟล์ Excel 2007
- สามารถแสดงข้อมูลได้หลายรูปแบบ เช่น scatter plot 3D
- สามารถแสดงผลโมเดลที่สวยงามและแก้ไขการแสดงผลให้สามารถอ่านได้ง่ายขึ้น
- สามารถบันทึกไฟล์โมเดลออกเป็นไฟล์ภาพประเภทต่างๆ เช่น PNG, JPG หรือ PDF
- มีวิธีการเตรียมข้อมูล (preprocess) และการวิเคราะห์ได้หลากหลายรูปแบบ

เมื่อเริ่มต้นใช้งาน RapidMiner Studio 7 จะแสดงหน้าต่างเริ่มต้นซึ่งประกอบด้วย 4 เมนูหลัก ดังนี้

LEARN เป็นหน้าที่รวบรวมและแสดงวิธีการใช้งานของ RapidMiner Studio 7 ซึ่งทำ Link ไปยังหน้าเว็บที่แสดงการใช้งานในรูปแบบ Document, VDO และมีบทเรียนฝึกหัดให้ทำตาม 3 บท คือ

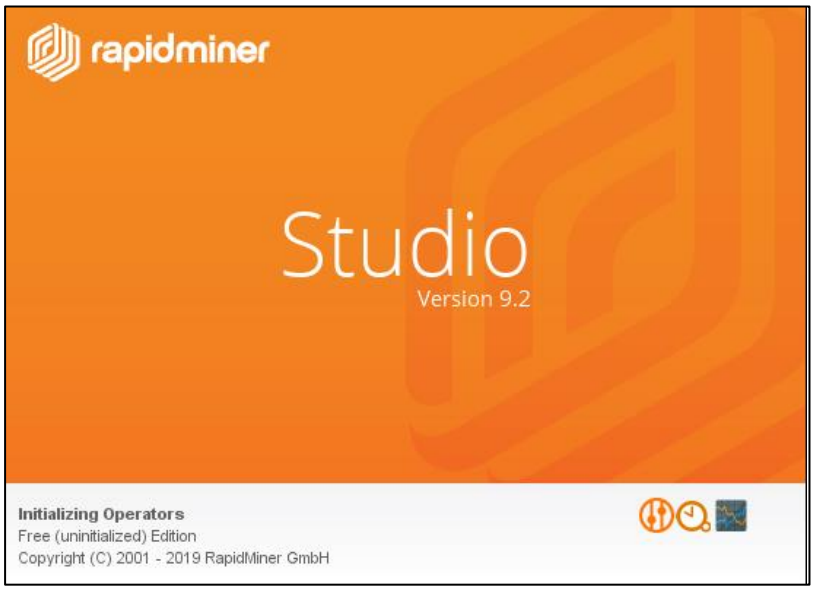
- Basic สอนพื้นฐานการใช้งาน
- Data Handling สอนการจัดการข้อมูล
- Modeling, Scoring and Validation โดยสอนเกี่ยวกับการสร้างโมเดลเพื่อทำ Prediction การวิเคราะห์ผลและนำผลลัพธ์ข้อมูลเชิงลึกที่คาดการณ์ได้มาใช้จริงและสอนกระบวนการยืนยันความถูกต้องของโมเดล

NEW PROCESS สร้างโปรเซสใหม่เพื่อเริ่มการใช้งาน RapidMiner ซึ่งทุกครั้งที่ต้องการสร้าง งานใหม่ที่แตกต่างจะต้องสร้างโปรเซสใหม่

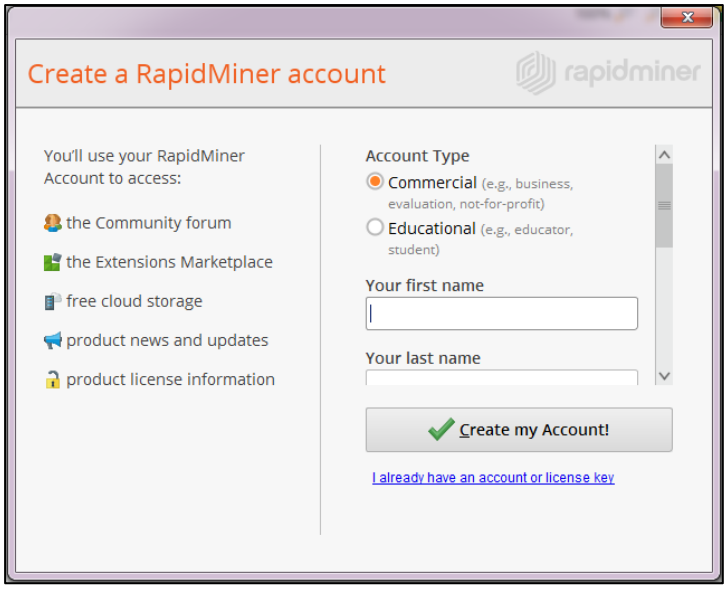
OPEN PROCESS เปิดโปรเซสเก่าที่เคยสร้างไว้เพื่อดูหรือแก้ไข โดยโปรเซสที่สร้างไว้แล้วสามารถ Reuse ได้ หรือ ส่งให้คนอื่นได้

เริ่มต้นใช้งาน RapidMiner Studio 9

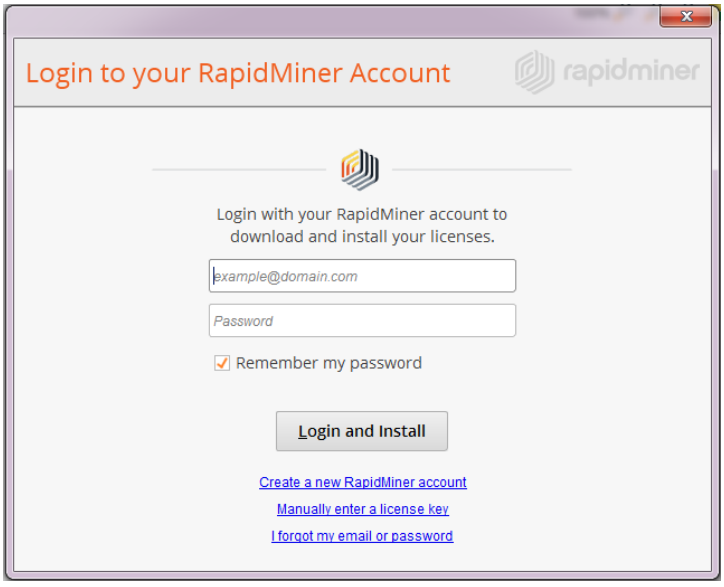
เมื่อเรา Download RapidMiner Studio 9 ให้คลิกใช้งานที่ Icon หลังจากนั้นจะขึ้นหน้าต่าง Welcome ของโปรแกรม RapidMiner Studio 9 ดังรูป



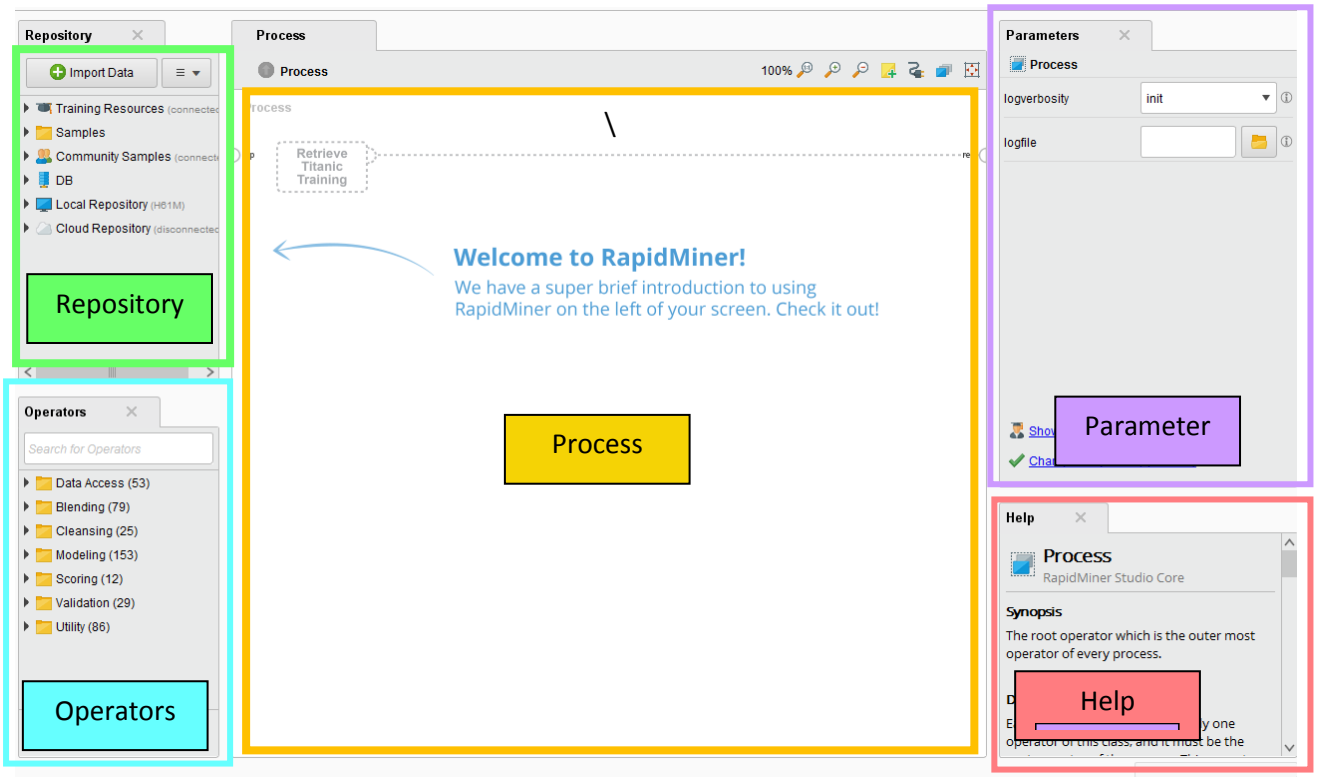
หลังจากเข้าสู่หน้าโปรแกรม RapidMiner Studio 9 จะให้เรา ลงทะเบียนเข้าสู่โปรแกรมซึ่งโปรแกรมนี้มีให้ใช้ทั้งใช้งานฟรี และชำระเงินเพื่อใช้งานบาง function หากใครมี account ของโปรแกรม RapidMiner Studio 9อยู่แล้ว สามารถเข้าใช้งานได้โดยใช้ account เดิมได้ หรือจะใส่ Key เพื่อใช้งานโดยเลือกดที่ I already an account or license key.



หลังจากกด I already an account or license key. เข้ามาแล้วสามารถ เข้าสู่ account โดยการกรอก E-mail และ Password และเข้าใชงานได้ทันที หรือเลือก ที่ปุ่ม Manually enter license Key . เพื่อใส่ license Key ที่ได้จากการซื้อโปรแกรมใส่เพื่อใช้งานได้เช่นกัน

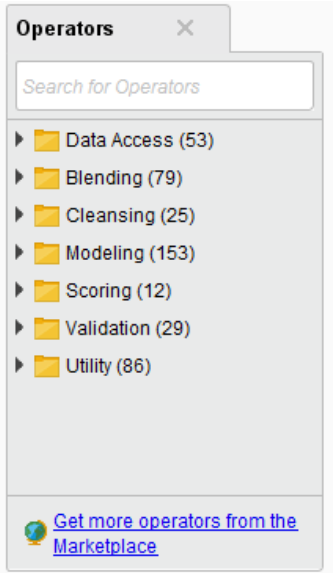


องค์ประกอบของ RapidMiner Studio 9



1. Repository เป็นส่วนสำหรับจดการไฟล์ RapidMiner จะจัดการข้อมูลจาก 3 แหล่ง คือ DB , Local (ในเครื่องคอมพิวเตอร์ที่ใช้อยู่), และ Cloud Repository โดยเก็บไฟล์ Data Set และ Process ต่าง ๆ แยกเก็บไว้คนละโฟลเดอร์

2.Operators เป็นส่วนที่ใช้เก็บตัวโอเปอเรเตอร์ ที่ใช้ในการท างานทั้งหมด ซึ่งจัดเป็นกลุ่ม ๆ โดยกลุ่มที่ใช้งานคล้ายคลึงกันจะจัดอยู่ในกลุ่มเดียวกัน มี 8 กลุ่ม คือ



2.1 Data Access

2.2 Blending

2.3 Cleansing

2.4 Modeling

2.5 Scoring

2.6 Validation

2.7 Utility

2.8 Extensions

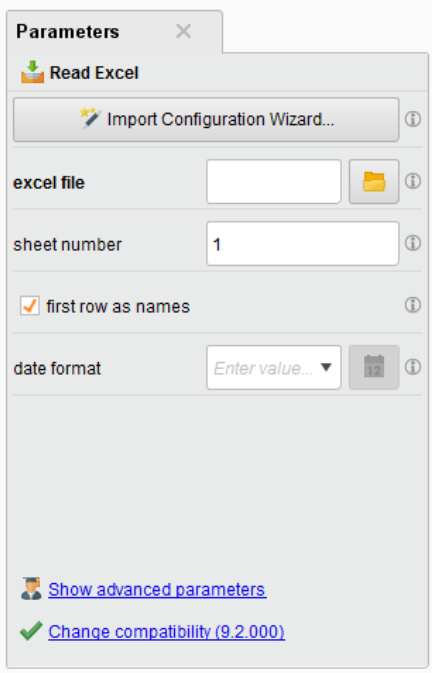
โอเปอเรเตอร์แต่ละตัวจะประกอบด้วย

- ชื่อของโอเปอเรเตอร์
- อินพุท พอร์ต (Input port) เป็นส่วนรับข้อมูลเข้ามาประมวลผล
- เอาท์พุต พอร์ต (Output port) เป็นส่วนส่งผลลัพธ์ที่ประมวลผลได้

โดยชื่ออินพุท พอร์ตและชื่อเอาท์พุต พอร์ต จะแสดงด้วยตัวอักษร 3 ตัวแรกของชื่อพอร์ต เช่น exa ย่อมาจาก example set เป็นต้น

3. Process เป็นหน้าหลักในการท างานในการสร้างโปรเซสสำหรับทำ Machine Learning ของซอฟต์แวร์นี้ โดยจะน าโอเปอเรเตอร์มาประกอบเพื่อสร้างโปรเซสขึ้นตามวัตถุประสงค์ของโจทย์ที่ตั้งไว้

4. Parameters เป็นส่วนสำหรับแสดงพารามิเตอร์ (parameter) ที่เกี่ยวข้องกับแต่ละโอเปอเรเตอร์ เช่น โอเปอเรเตอร์ Read Excel ที่ใช้เพื่ออ่านไฟล์ประเภท Excel จะมีพารามิเตอร์ที่เกี่ยวข้อง เช่น ชื่อและที่อยู่ของ ไฟล์ Excel เป็นต้น



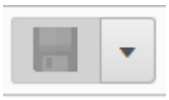
แสดงส่วนประกอบพารามิเตอร์ของโอเปอเรเตอร์ Read Excel

5. Help เป็นส่วนช่วยเหลือ ซึ่งจะแสดงรายละเอียดของตัวโอเปอเรเตอร์ที่เลือกใช้งานอยู่ ส่วนช่วยเหลือของ RapidMiner จะบอกเพียงหน้าที่และรายละเอียดคราว ๆ ของโอเปอเรเตอร์ หากต้องการดูรายละเอียดมากกว่านั้นต้องไปที่ Jum to Tutorai Process ซึ่งจะลิงก์ไปยังเว็บไซต์ที่มีรายละเอียดที่เกี่ยวกับโอเปอเรเตอร์ที่ใช้อยู่

นอกจากทั้ง 5 ส่วนใหญ่ ๆ ที่ได้อธิบายแล้วยังมีส่วนเมนูด้านบนเพิ่มเติมดังนี้

เมนูสำหรับสร้าง โพรเซสใหม่

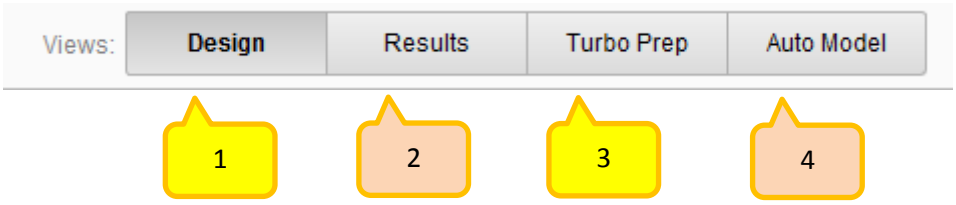
เมนูสำหรับการโหลดไฟล์ต่าง ๆ จาก repository



เมนูสำหรับการบันทึก โพรเซส หรือบันทึก โพรเซสเป็นชื่อใหม่

เมนูสำหรับสั่งให้ โพรเซสทำงาน

เมนูสำหรับยกเลิกการทำงาน โพรเซส



1. แสดงหน้าจอการออกแบบ Design
2. แสดงหน้าจอผลลัพธ์การทำงาน Results
3. แสดงหน้าต่าง ออกแบบมาเพื่อให้การเตรียมข้อมูลง่ายขึ้น
4. แสดงหน้าต่างการสร้างโมเดล อัตโนมัติ Auto Model

ตัวอย่างการสร้างโมเดล Decision Tree

ข้อมูลที่มีอยู่ในตารางต่าง ๆ ที่ประกอบไปด้วยแถวและคอลัมน์ ซึ่งจะเรียกในแถวเป็น ตัวอย่าง (Example) ส่วนคอลัมน์เรียก แอตทริบิวต์(Attribute) มีหน้าที่ 3 แบบ คือ

1. ID เป็นแอตทริบิวต์ที่แสดงหมายเลขของข้อมูลหรือ primary key ในฐานข้อมูล

2. แอททริบิวต์ทั่วไป (Attribute) เป็นแอตทริบิวต์ปกติที่จะใช้ในการสร้างโมเดลหรือเรียกว่าฟีเจอร์ (feature) หรือตัวแปรต้น(independent variable)

3. Label คือ แอตทริบิวต์ที่เป็นค าตอบที่เราต้องการจะสร้างโมเดลขึ้นมาทำนาย หรือ เรียกว่า คลาส (class) หรือตัวแปรตาม (dependent variable)

Row No.	รหัสนักศึกษา	endgrade	เพศ	กรุ๊ปเลือด	น้ำหนัก	ส่วนสูง	สถานภาพกา...	std_
1	5822041025	2.210	ญ	O	46	159	DRP	ปกติ
2	5122010235	2.490	ญ	O	55	163	END	ปกติ
3	4922010357	2.280	ญ	O	48	162	END	ปกติ
4	5122010225	2.500	ญ	O	45	156	END	ปกติ
5	4922010315	2.820	ญ	O	47	156	END	ปกติ
6	4922010336	2.400	ญ	O	50	170	END	ปกติ
7	5022010174	2.240	ญ	O	49	159	END	ปกติ
8	4922010321	2.310	ญ	O	50	153	END	ปกติ
9	4922010353	2.530	ญ	O	40	153	END	ปกติ
10	5022010208	2.380	ญ	O	48	158	END	ปกติ
11	4922010278	2.370	ญ	O	61	162	END	ปกติ
12	4922010343	2.560	ญ	O	46	147	END	ปกติ

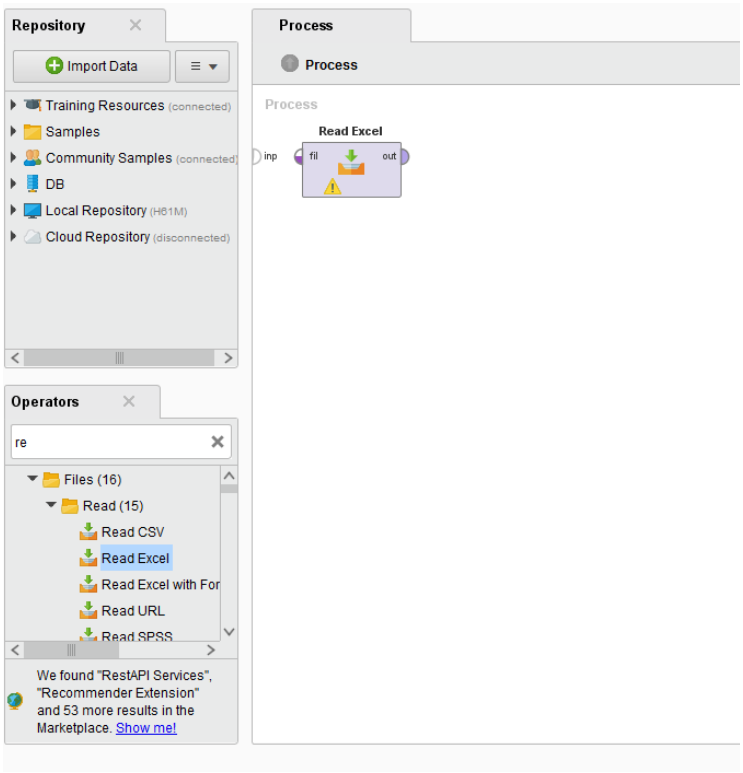
ID **Label** **Attribute**

ประเภทของข้อมูลที่เก็บไว้ในแต่ละแอตทริบิวต์ มีดังนี้

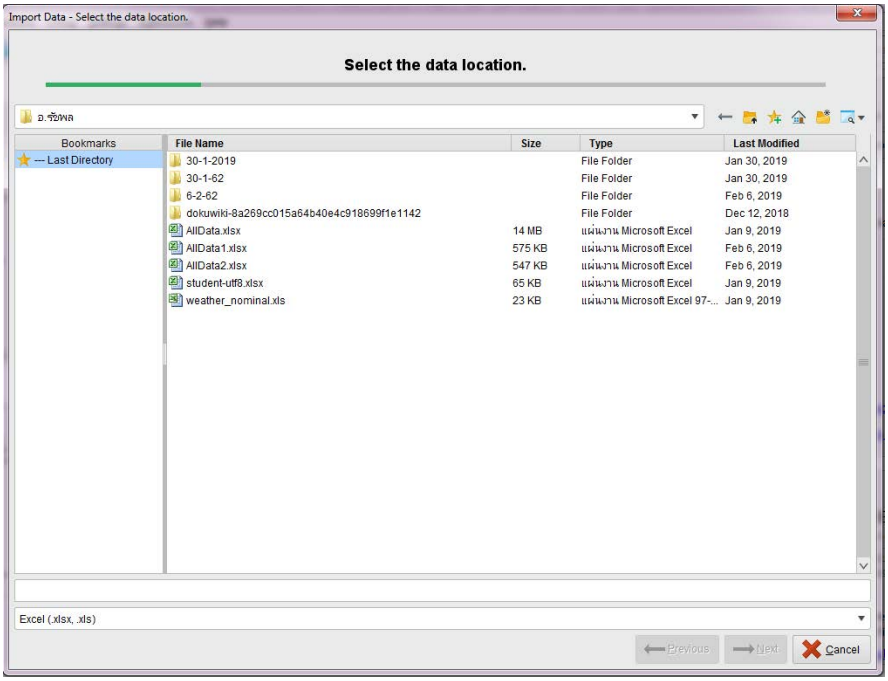
1. Polynominal เป็นข้อมูลประเภท Category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าแตกต่างกันมากกว่า 2 ค่า
2. Binominal เป็นข้อมูลประเภท Category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าเพียง 2 ค่าเท่านั้น
3. Numeric หรือ Integer ข้อมูลประเภทตัวเลข
4. Text ข้อมูลประเภทข้อความ

ขั้นตอนการสร้าง Decision Tree

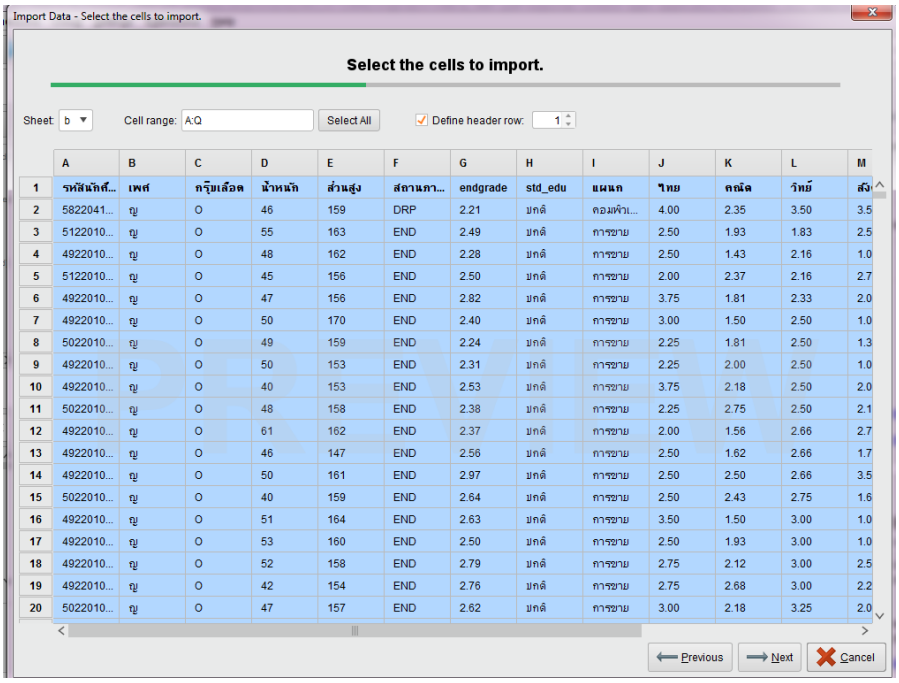
1. ไปที่ Operators>Data Access>Files>Read ลากโอเปอเรเตอร์ชื่อ Read Excel มาวางที่ Process (สามารถพิมพ์คำว่า Read ในช่อง Search for Operators เพื่อค้นหาโอเปอเรเตอร์ Read Excel ก็ได้) จากนั้น ลากเส้นเชื่อมจากพอร์ตที่ชื่อout (output) ของโอเปอเรเตอร์ Read Excel ไปยังพอร์ตที่ชื่อว่า res(result)



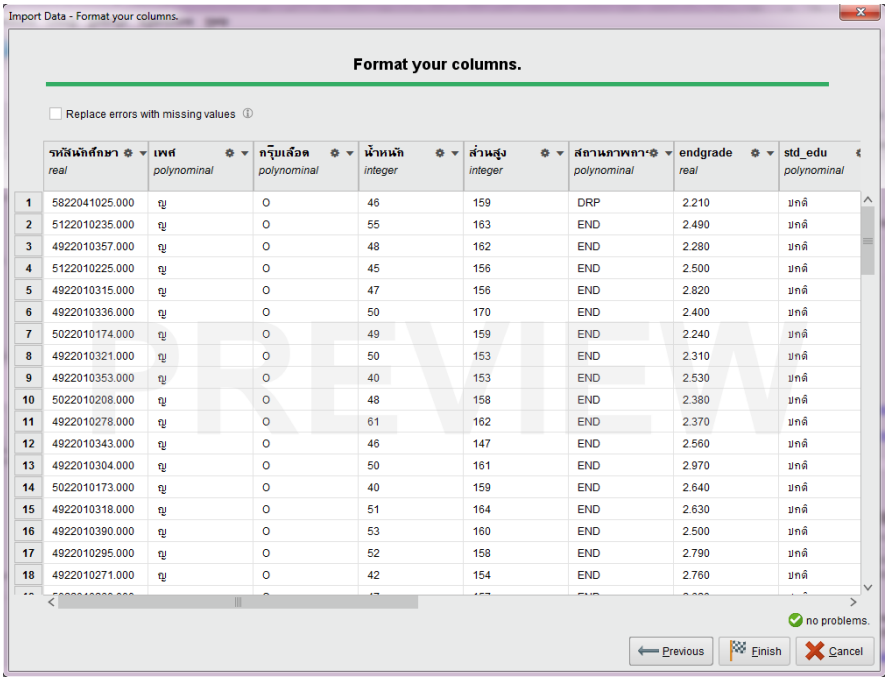
2. ในส่วนค่าพารามิเตอร์คลิกเลือก Import Configuration Wizard จะปรากฏหน้าต่างดังภาพ ให้เลือก ข้อมูลซึ่งเป็นไฟล์ประเภท excel จากนั้นคลิก next



3. หน้านี้จะเป็นการแสดงการเลือกไฟล์ข้อมูลหรือ Data Set ที่จะนำมาใช้งาน โดยไฟล์ที่เราเลือก Read เป็นไฟล์ excel ดังนั้นไฟล์ที่เราเปิดได้จะเป็นไฟล์ excel



4. หน้านี้เป็นหน้าต่างที่เราสามารถเลือกได้ว่าจะนะ Attribute ใดบ้างใช้งานได้บ้างหลังจากเราเลือก Attribute แล้วก็กด Next

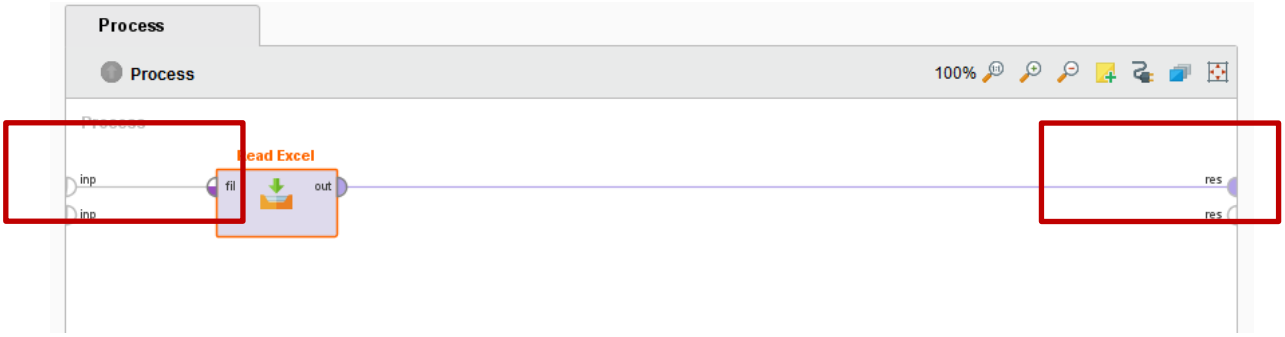


5. หน้าถัดมาให้ทำการเปลี่ยนที่ชื่อ Attribute ให้เป็น label เพื่อใช้ในการทำนายผลของเรา ในที่นี้เรากำหนด endgrade เป็น label และกำหนด รหัสนักศึกษา เป็น ID เพื่อเป็น Key ในข้อมูล จากนั้นคลิก Finish

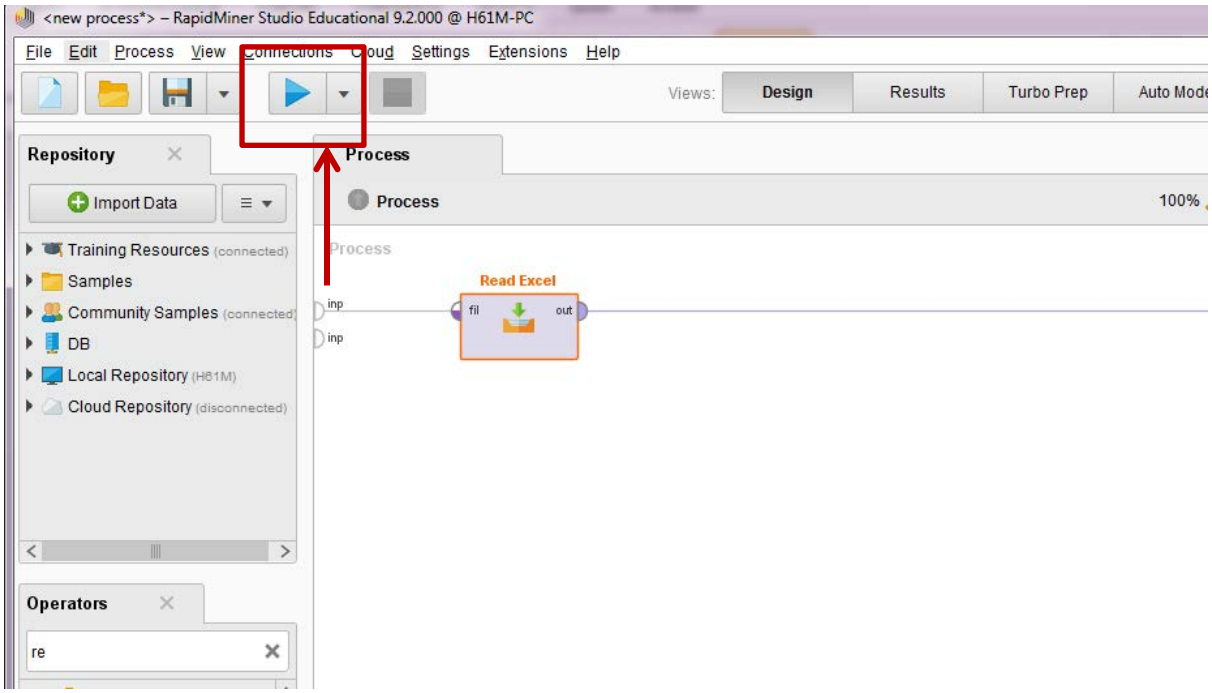
	รหัสนักศึกษา real id	เพศ polynominal	กรุ๊ปเลือด polynominal	น้ำหนัก integer	ส่วนสูง integer	สถานภาพการ polynominal	endgrade real label	std_edu polynominal
1	5822041025.000	ญ	O	46	159	DRP	2.210	ปกติ
2	5122010235.000	ญ	O	55	163	END	2.490	ปกติ
3	4922010357.000	ญ	O	48	162	END	2.280	ปกติ
4	5122010225.000	ญ	O	45	156	END	2.500	ปกติ
5	4922010315.000	ญ	O	47	156	END	2.820	ปกติ
6	4922010336.000	ญ	O	50	170	END	2.400	ปกติ
7	5022010174.000	ญ	O	49	159	END	2.240	ปกติ
8	4922010321.000	ญ	O	50	153	END	2.310	ปกติ
9	4922010353.000	ญ	O	40	153	END	2.530	ปกติ
10	5022010208.000	ญ	O	48	158	END	2.380	ปกติ
11	4922010278.000	ญ	O	61	162	END	2.370	ปกติ
12	4922010343.000	ญ	O	46	147	END	2.560	ปกติ

ID **Label**

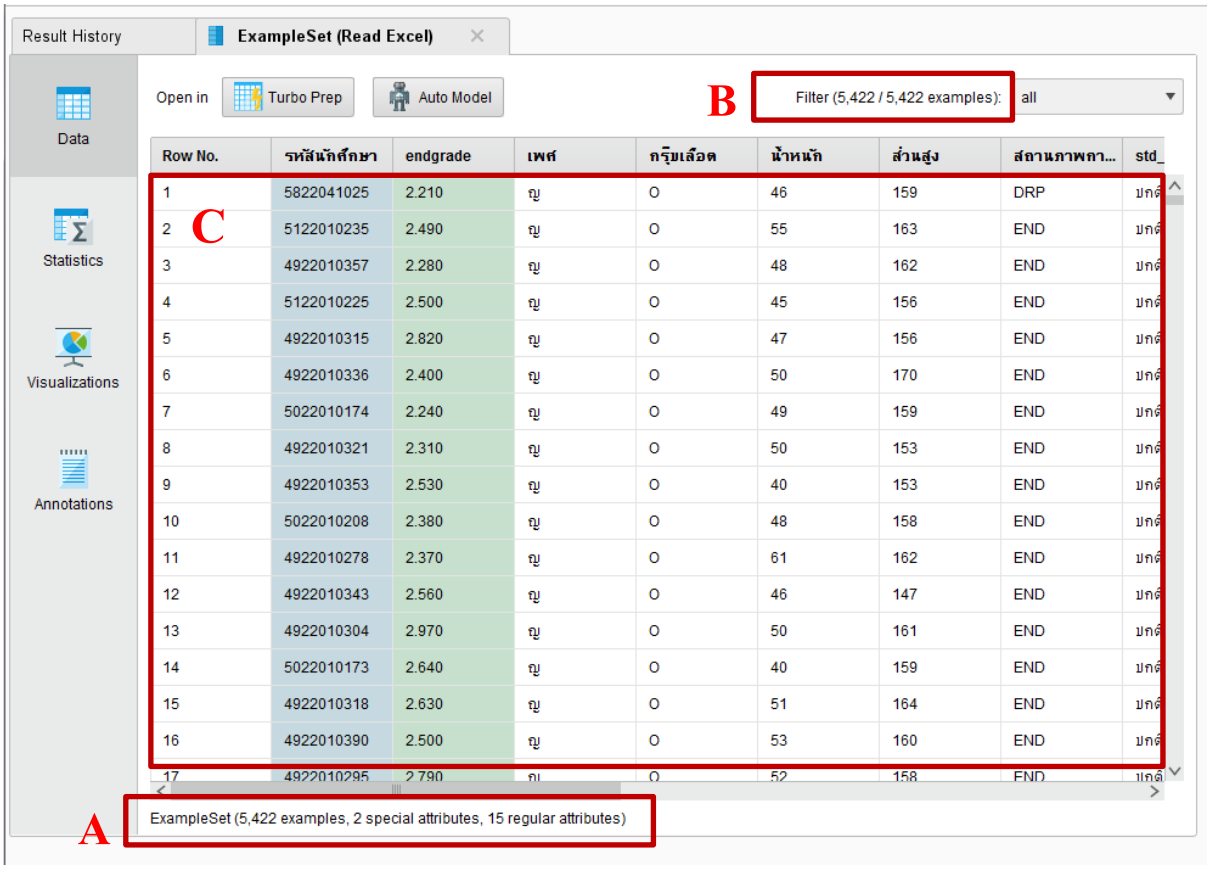
6. หลังจากเลือไฟล์และกด Finish จะขึ้นหน้าต่างการทำงาน หลังจากนั้นลากเส้นเชื่อมการทำงานให้ได้ดังรูป



7. หลังจากเราลากเส้นเชื่อมการทำงานแล้ว ให้เรา คลิก Run Process จะปรากฏหน้าต่างผลลัพธ์ของการทำงาน



8. เมื่อเราสั่งโปรแกรมทำงานแล้ว มีแอตทริบิวต์พิเศษจำนวน 1 แอตทริบิวต์ (ตัวที่เราเลือกเป็น labelจะมีแถบสีเป็นสีเขียว) และแอตทริบิวต์ทั่ว ๆ ไป ที่ใช้สำหรับสร้างโมเดล ดังภาพ



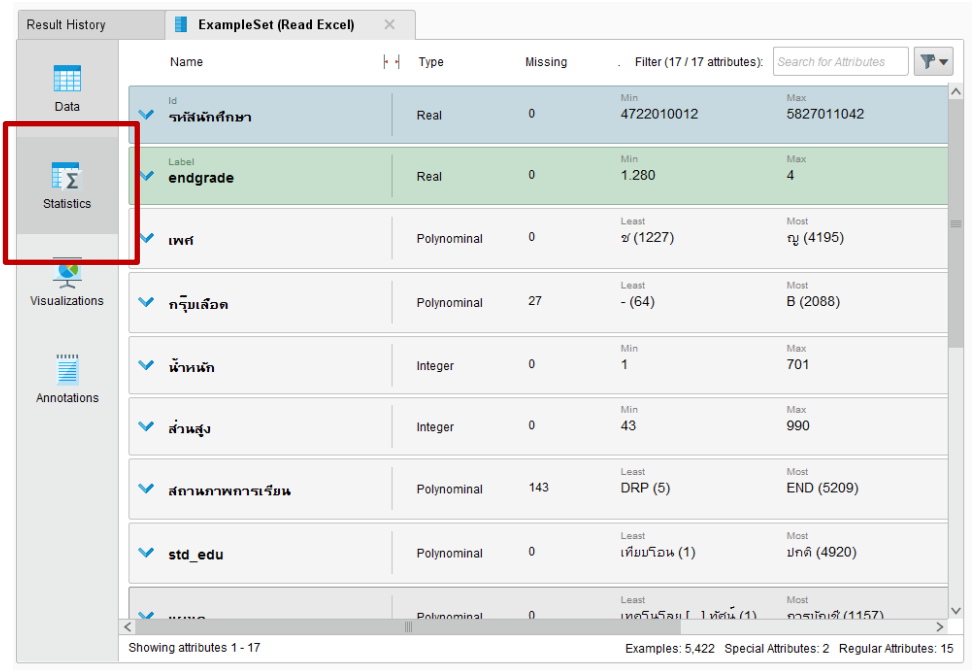
Row No.	รหัสนักศึกษา	endgrade	เพศ	กรุ๊ปเลือด	น้ำหนัก	ส่วนสูง	สถานภาพครา...	std_
1	5822041025	2.210	ญ	O	46	159	DRP	ปกติ
2	5122010235	2.490	ญ	O	55	163	END	ปกติ
3	4922010357	2.280	ญ	O	48	162	END	ปกติ
4	5122010225	2.500	ญ	O	45	156	END	ปกติ
5	4922010315	2.820	ญ	O	47	156	END	ปกติ
6	4922010336	2.400	ญ	O	50	170	END	ปกติ
7	5022010174	2.240	ญ	O	49	159	END	ปกติ
8	4922010321	2.310	ญ	O	50	153	END	ปกติ
9	4922010353	2.530	ญ	O	40	153	END	ปกติ
10	5022010208	2.380	ญ	O	48	158	END	ปกติ
11	4922010278	2.370	ญ	O	61	162	END	ปกติ
12	4922010343	2.560	ญ	O	46	147	END	ปกติ
13	4922010304	2.970	ญ	O	50	161	END	ปกติ
14	5022010173	2.640	ญ	O	40	159	END	ปกติ
15	4922010318	2.630	ญ	O	51	164	END	ปกติ
16	4922010390	2.500	ญ	O	53	160	END	ปกติ
17	4922010295	2.790	ญ	O	52	158	END	ปกติ

A แสดงจำนวนตัวอย่างและแอตทริบิวต์ที่ปรากฏในข้อมูลซึ่งในไฟล์ตัวอย่างนี้มีจำนวน 5,422 ตัวอย่าง 1 แอตทริบิวต์ประเภทลาเบล และ 1 แอตทริบิวต์ประเภท ไอดี และ 15 แอตทริบิวต์ทั่วไป

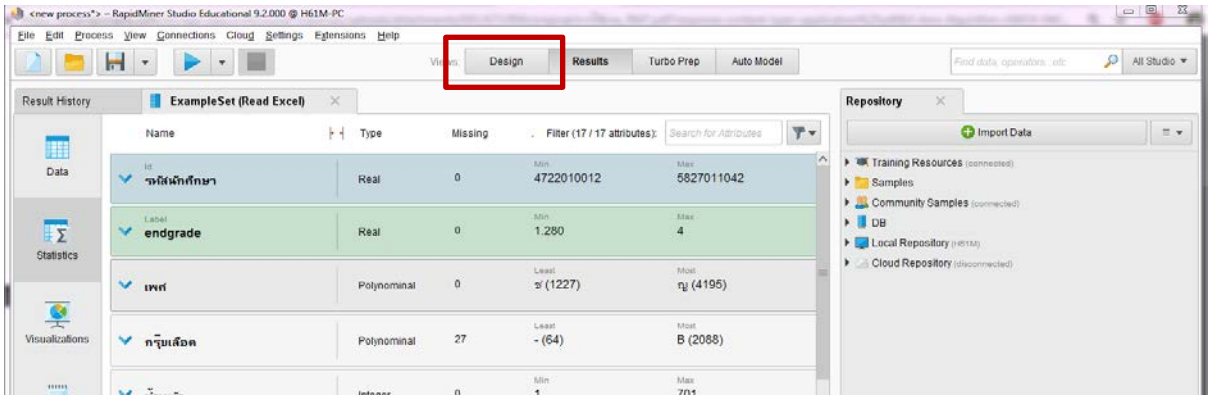
B ส่วนของการกรองข้อมูล (filter) ซึ่งมีให้เลือกได้ว่าจะดูข้อมูลทั้งหมดหรือข้อมูลที่มีความผิดพลาด (missing_attributes) อยู่

C ในส่วนของตารางเราสามารถคลิกเลือกที่ชื่อแอตทริบิวต์เพื่อทำการเรียงลำดับข้อมูลได้ โดย ตารางข้อมูลจะแบ่งแอตทริบิวต์ออกเป็น 3 แบบคือ

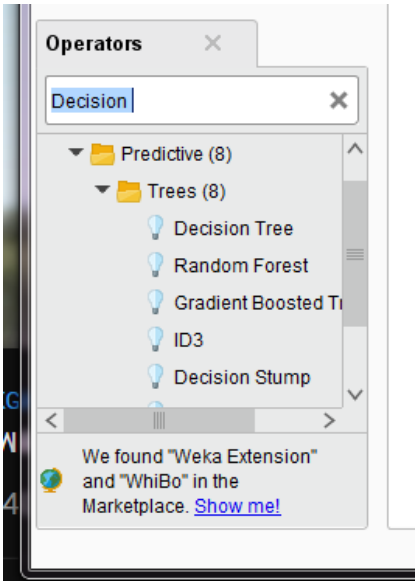
- แอตทริบิวต์ที่เป็นลาเบลแสดงด้วยคอลัมน์สีเขียว
- แอตทริบิวต์ที่เป็นไอดีแสดงคอลัมน์สีฟ้า
- แอตทริบิวต์ทั่วไปแสดงด้วยคอลัมน์ที่เป็นสีเทา



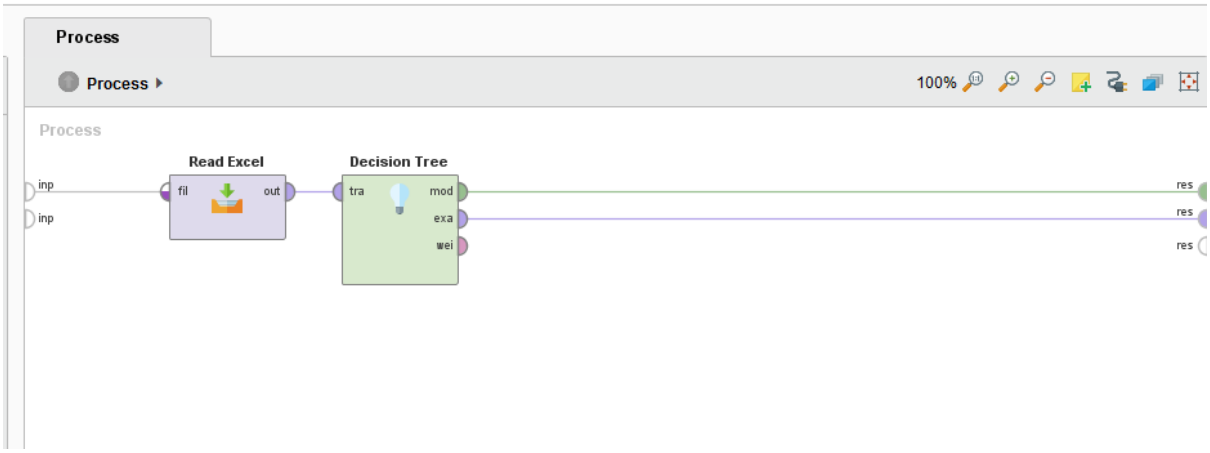
9. คลิกเลือก Statistics ด้านซ้ายมือ เพื่อแสดงค่าสรุปทางสถิติของแอตทริบิวต์ต่าง ๆ โดยจะแสดงชื่อประเภทของข้อมูลที่เก็บอยู่ กราฟแสดงค่าความถี่ของข้อมูลในแต่ละแอตทริบิวต์ดังภาพ



10. คลิกเลือกมุมมอง Design ต่อไปเราจะทำการสร้างโมเดล Decision Tree โดยการเลือกโอเปอเรเตอร์ Decision Tree จากส่วนของ Operators โดยการพิมพ์ตรงช่องค้นหา โดยพิมพ์คำว่า Decision กดปุ่ม Enter ก็จะปรากฏโอเปอเรเตอร์ Decision Tree ขึ้นมา หรือจะทำการเลือกจากหมวด Modeling >> Classification and Regression >> Tree Induction



11. ลากโอเปอเรเตอร์ Decision Tree มาวางในส่วนของ Process ตรงเส้นที่เชื่อมต่อเดิมที่โอเปอเรเตอร์ Read Excel ลากไว้ (โปรแกรมจะทำการเชื่อมโอเปอเรเตอร์ทั้งสองตัวทันทีจากพอร์ต out ของโอเปอเรเตอร์ Read Excel ไปยังพอร์ต tra (training) ของโอเปอเรเตอร์ Decision Tree เพื่อเป็นการส่งข้อมูลไปสร้างโมเดล

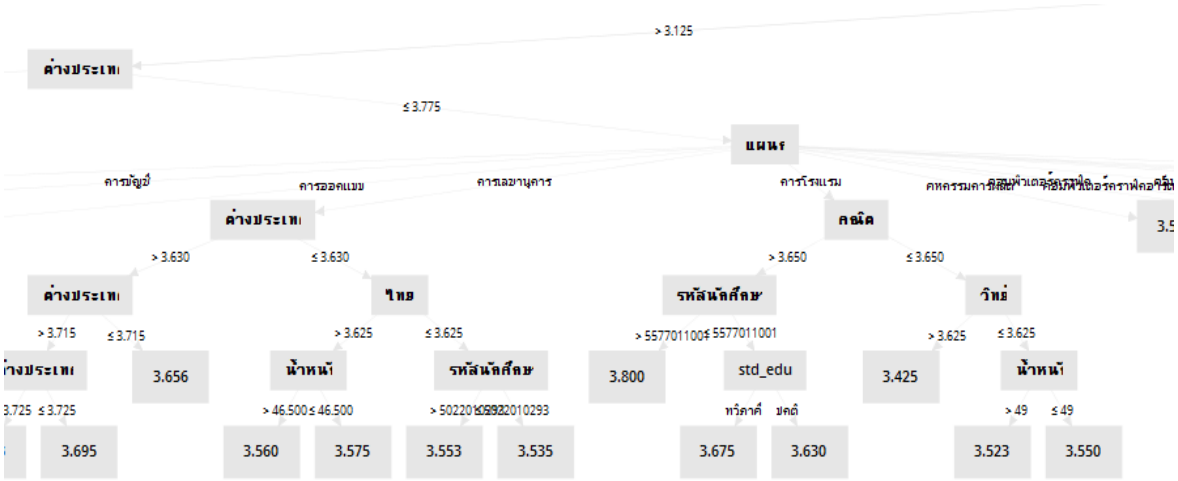


12. ลากเส้นเชื่อมจากพอร์ต mod (model) และพอร์ต exa (example) ของโอเปอเรเตอร์ Decision Tree ไปยังพอร์ต res (result) ทั้งสองพอร์ต เพื่อไปแสดงในส่วนของหน้าจอผลลัพธ์โดยพอร์ต mod จะทำการส่งโมเดล Decision Tree ที่สร้างออกไปแสดงในรูปต้นไม้ และพอร์ต exa จะส่งข้อมูลที่ import เข้ามาไปแสดงในรูปแบบตาราง

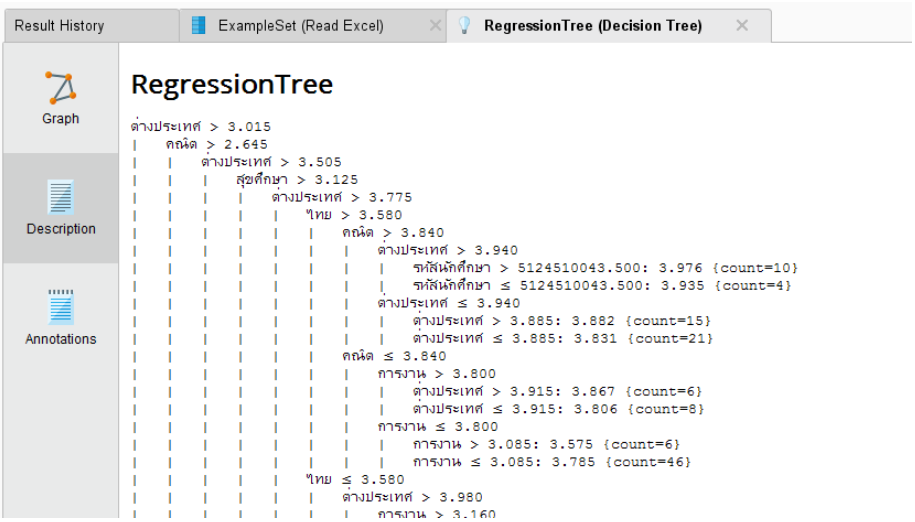
13. จากนั้นคลิก Run Process จะได้รูปโมเดลต้นไม้ ซึ่งโมเดลต้นไม้ที่สร้างได้มีส่วนประกอบสำคัญ 3 ส่วน คือ

- ในโมเดล Decision Tree จะมีโหนดต่าง ๆ 2 ประเภท คือ
  - โหนดที่เป็นแอตทริบิวต์แสดงด้วยรูปสี่เหลี่ยมที่มีมุมโค้ง
  - โหนดลาเบลแสดงด้วยรูปสี่เหลี่ยมที่มีกราฟแสดงสีต่าง ๆ อยู่ด้วย ในตัวอย่างนี้มีหลาย label แต่หากมีการกำหนดเกณฑ์มาตรฐานของคะแนนในตัวอย่างไล่เป็นระดับตามเกรด ตัว label ก็จะมีคำตอบตามลำกับที่เราตั้งค่าไว้ จะมีกราฟสีน้ำเงิน

- ส่วนของ Zoom ใช้สำหรับย่อขยายรูปโมเดล
- ส่วนของ Mode จะใช้สำหรับปรับโหมดของการใช้งานเมาส์

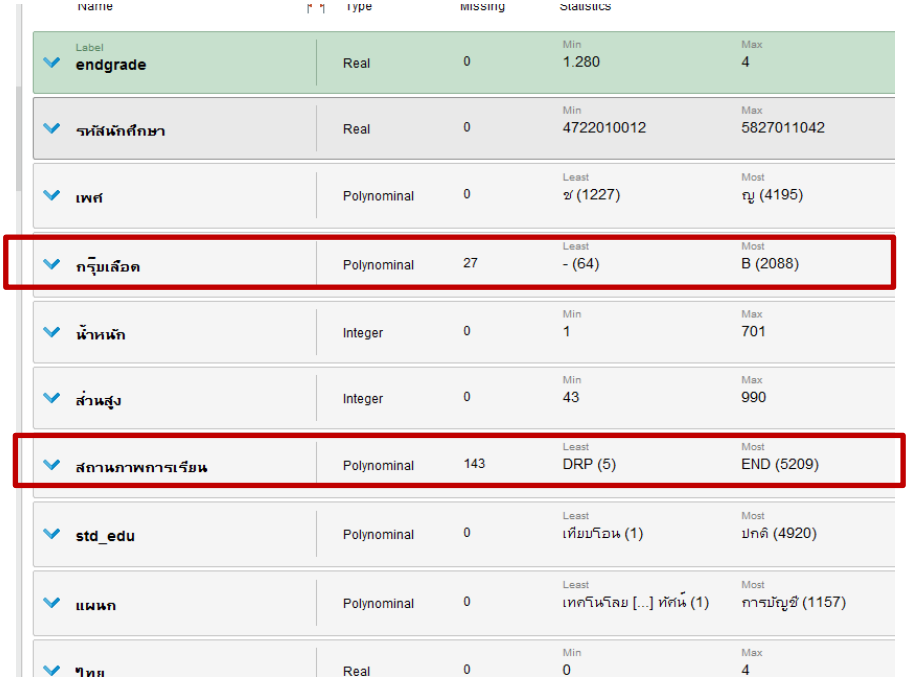


14. ในหน้าต่าง Description จะเป็นโค๊ดข้อความที่เราสามารถนำมาเขียนโปรแกรมเพื่อใช้ในการทำนายได้

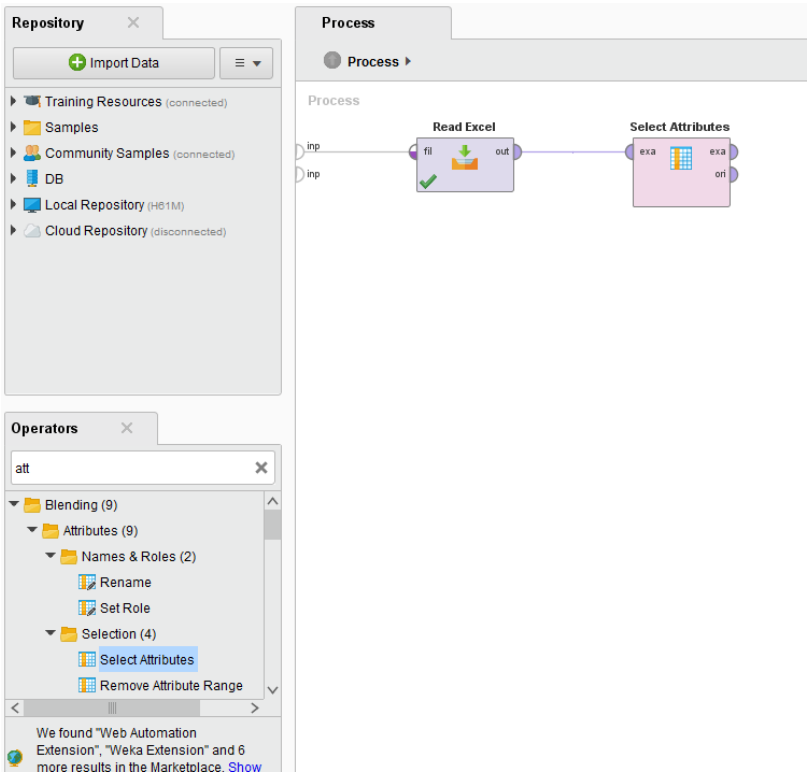


การจัดการข้อมูล (Data Manipulation)

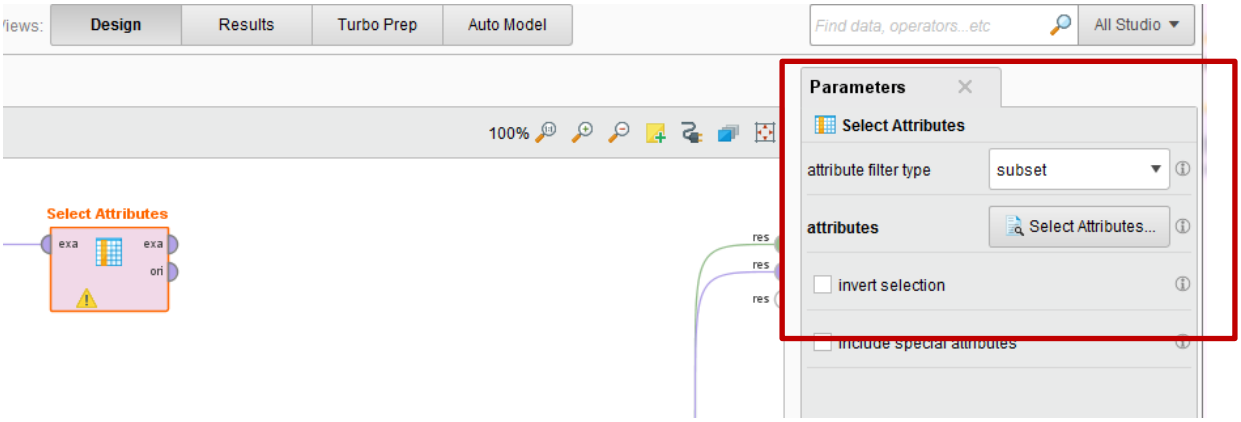
หลังจากที่เราลองทำ Decision Tree แล้วเราจำนำโมเดที่เราสร้างมาทำการทำนาย แต่เราจะทำนายไม่ได้หากยังมี Missng รวมถึงเราต้องเลือกใช้เพียงแอตทริบิวต์บางตัวที่จำมาใช่ในการทำนายเท่านั้น ในที่นี้ยังมี แอตทริบิวต์บางตัว ที่ติด Missng อยู่ดังนั้นเราจะนำตัวที่ติด Missng ออกเนื่องจากกว่าตัวที่เรานำออกนั้นไม่ได้มีผลต่อการทำนาย แต่หากว่าแอตทริบิวต์ที่ติด Missng มีผลต่อการทำนายเราอาจจะใช้เป็นค่าเฉลี่ย หรือการแทนค่าเข้าไปแทน



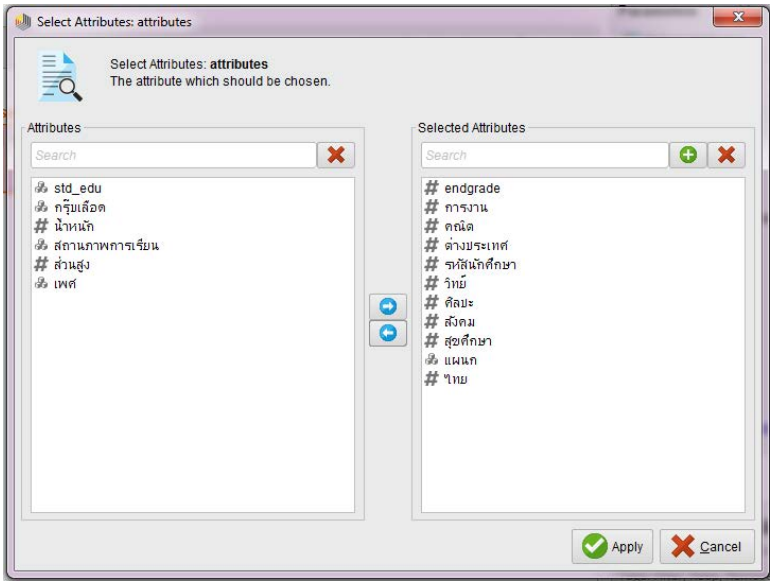
1. พิมพ์ค้นหาคำว่า Select Attributes แล้วลาก Operators มาวางที่หน่างการทำงาน และลากเส้นเชื่อมระหว่าง Out ของ Excel มาที่ exa ของ Select Attributes



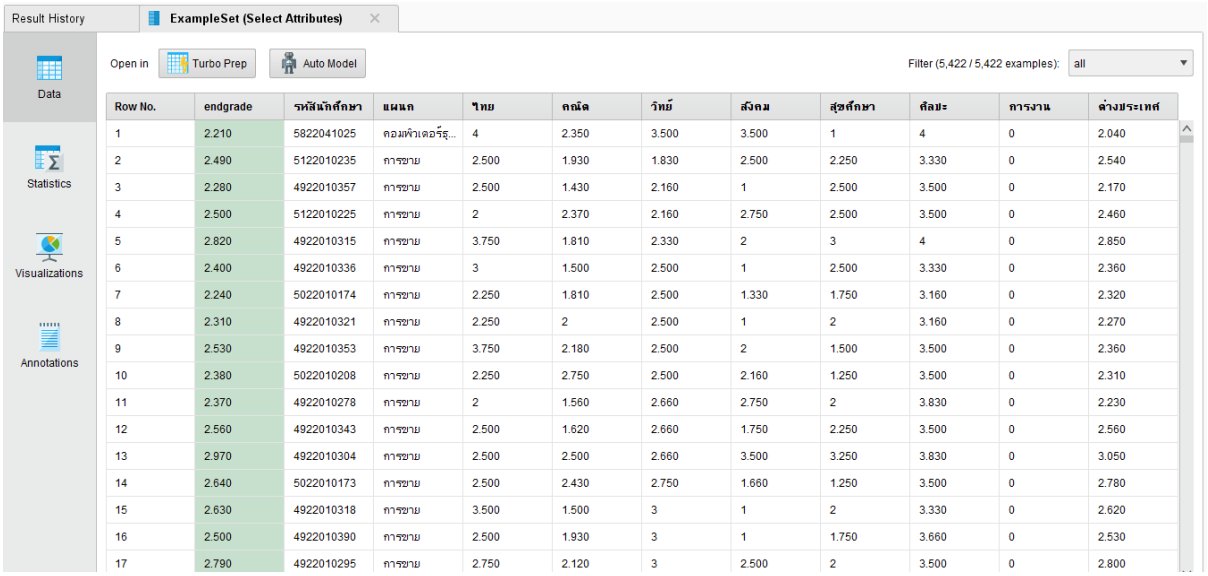
2. เลือกที่ Parameters ของ Select Attributes แล้วเลือก Attributes filter type เป็น subset หลังจากนั้นเลือก Select Attributes



3.หลังจากที่เราเลือก Select Attributes จะมีหน้าต่างขึ้นมาให้เราเลือก Attributes ที่เราต้องการที่จำใช้ในการทำนาย เมื่อเลือกข้อมูลที่ต้องการได้กด Apply และ สั่งรันโปรแกรมให้โปรแกรมทำงาน



4. หลังจากรันการทำงานของโปรแกรมแล้ว จะแสดงเฉพาะ Attributes ที่เราเลือกเท่านั้นเพื่อนำมาใช้งาน หากว่าเราต้องการทำนายแล้วยังไม่ได้กำหนด label สามารถกำหนด label โดยกาพิมพ์ค้นหา Set Row ได้ในช่อง Operation

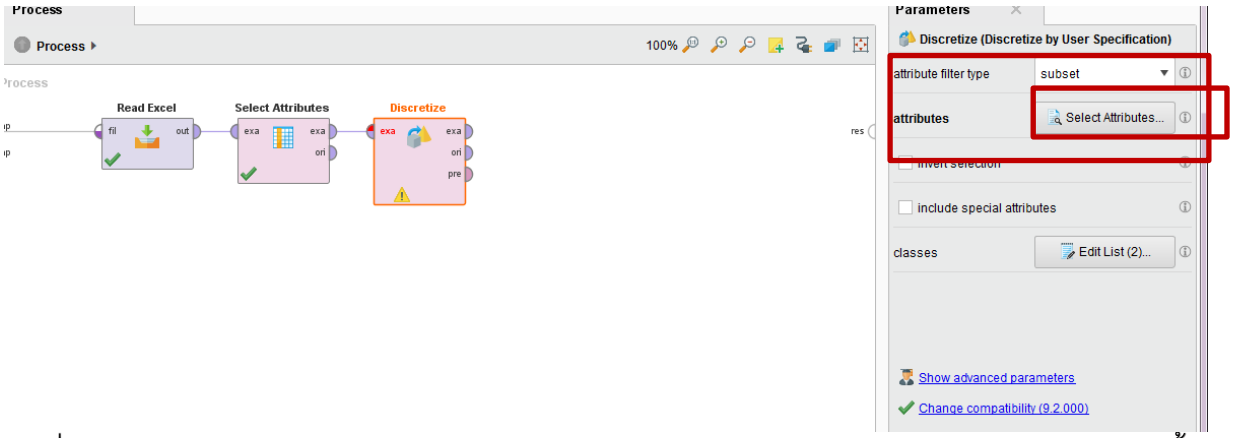


Row No.	endgrade	รหัสนักศึกษา	แผนก	ไทย	คณิต	วิทย์	สังคม	สุขศึกษา	ศิลปะ	การงาน	ต่างประเทศ
1	2.210	5822041025	คอมพิวเตอร์ธุ...	4	2.350	3.500	3.500	1	4	0	2.040
2	2.490	5122010235	การขาย	2.500	1.930	1.830	2.500	2.250	3.330	0	2.540
3	2.280	4922010357	การขาย	2.500	1.430	2.160	1	2.500	3.500	0	2.170
4	2.500	5122010225	การขาย	2	2.370	2.160	2.750	2.500	3.500	0	2.460
5	2.820	4922010315	การขาย	3.750	1.810	2.330	2	3	4	0	2.850
6	2.400	4922010336	การขาย	3	1.500	2.500	1	2.500	3.330	0	2.360
7	2.240	5022010174	การขาย	2.250	1.810	2.500	1.330	1.750	3.160	0	2.320
8	2.310	4922010321	การขาย	2.250	2	2.500	1	2	3.160	0	2.270
9	2.530	4922010353	การขาย	3.750	2.180	2.500	2	1.500	3.500	0	2.360
10	2.380	5022010208	การขาย	2.250	2.750	2.500	2.160	1.250	3.500	0	2.310
11	2.370	4922010278	การขาย	2	1.560	2.660	2.750	2	3.830	0	2.230
12	2.560	4922010343	การขาย	2.500	1.620	2.660	1.750	2.250	3.500	0	2.560
13	2.970	4922010304	การขาย	2.500	2.500	2.660	3.500	3.250	3.830	0	3.050
14	2.640	5022010173	การขาย	2.500	2.430	2.750	1.660	1.250	3.500	0	2.780
15	2.630	4922010318	การขาย	3.500	1.500	3	1	2	3.330	0	2.620
16	2.500	4922010390	การขาย	2.500	1.930	3	1	1.750	3.660	0	2.530
17	2.790	4922010295	การขาย	2.750	2.120	3	2.500	2	3.500	0	2.800

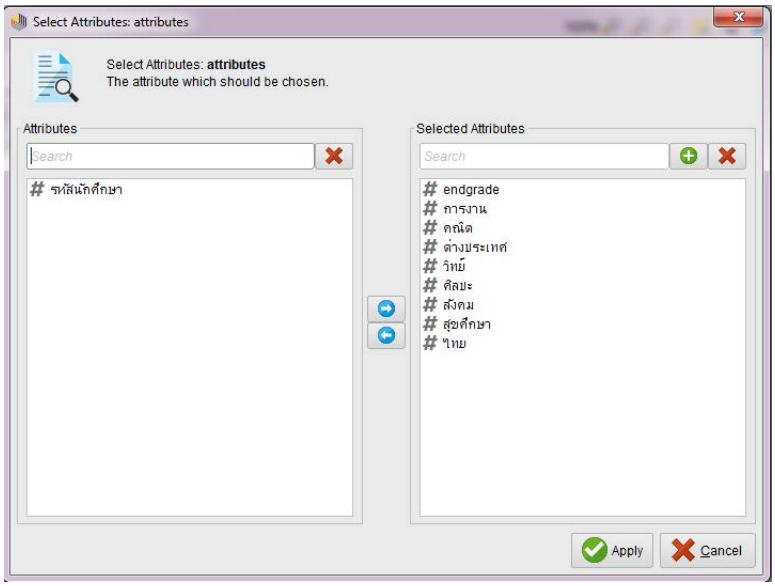
5. หลังจากได้ข้อมูลเบื้อต้นแล้ว ให้กำหนด ค่าของตัวเลขเป็นข้อความแบ่งระดับ โดยจะแบ่งระดับ เป็น

- ดีมาก = 4
- ปานกลาง = 3
- ต่ำ =2

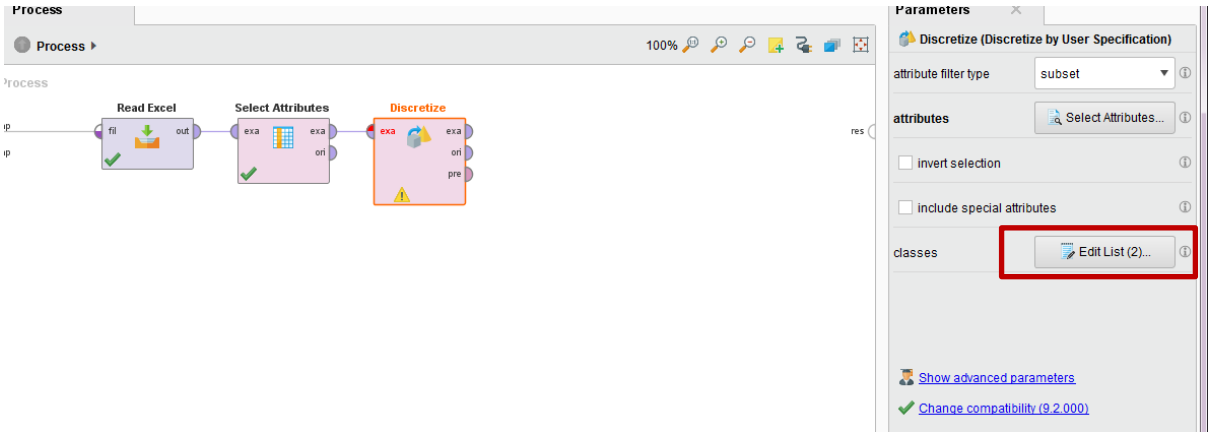
การตั้งค่าเพื่อให้ง่ายต่อการใช้งานโดยการ คลิกขวาเลือก Insert Operator >> Cleansing >> Binning >> Discretize by User Specification



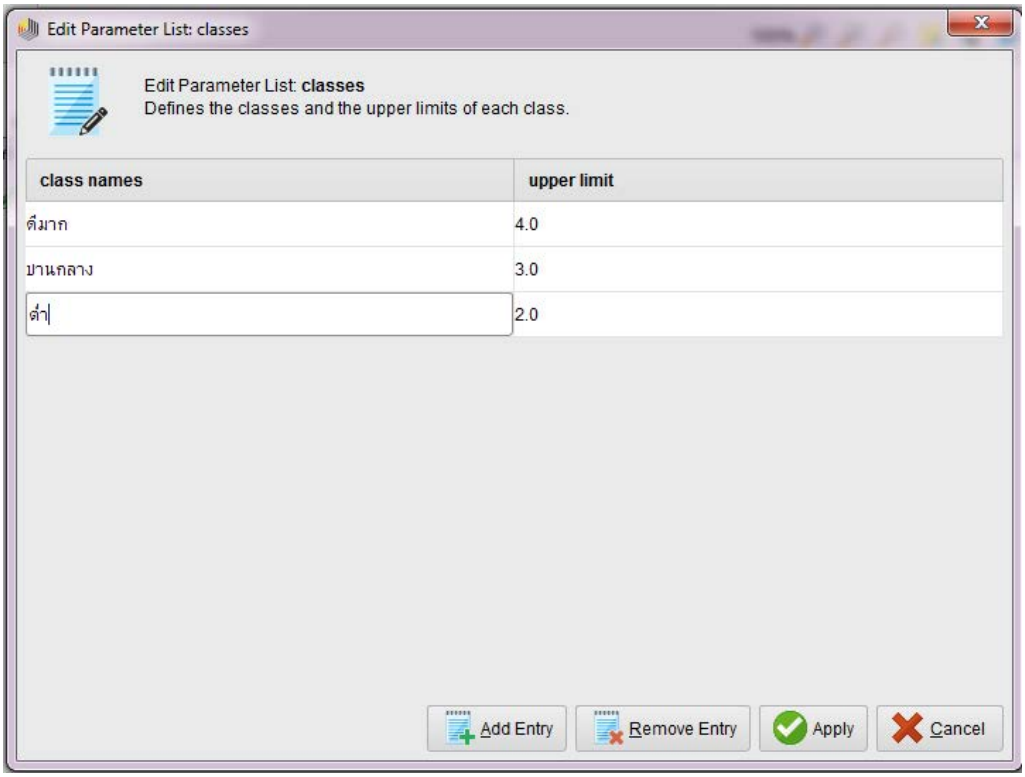
6. เมื่อได้ Operator แล้วเลือกตัว Operator แล้วเลือก Attributes filter type เป็น subset หลังจากนั้นเลือก Select Attributes เพื่อเลือก Attributes ที่ต้องการแทนที่ค่า เมื่อเลือก Attributes ที่ต้องการเสร็จแล้วกด apply



7. หลังจากนั้นกำหนดเกณฑ์ที่ต้องการเทียบกับคะแนน เพื่อแปลงเป็นระดับที่เราตั้งไว้ ในช่อง Classes เลือก Edit List



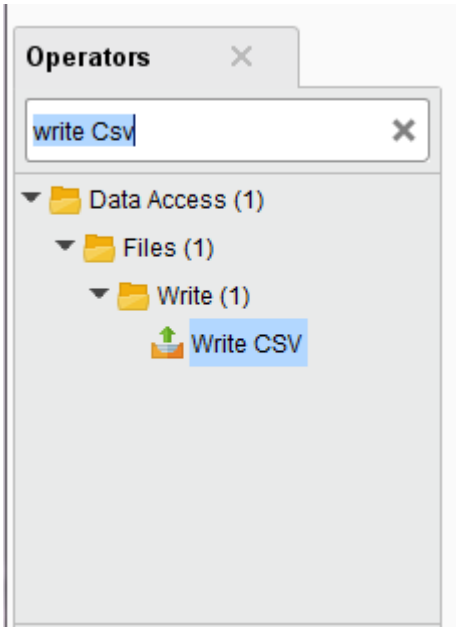
8. เมื่อกดเข้าไปจะมีหน้าต่าง ให้เรากำหนดเกณฑ์ เมื่อเรากำหนดเกณฑ์แล้วกด Apply



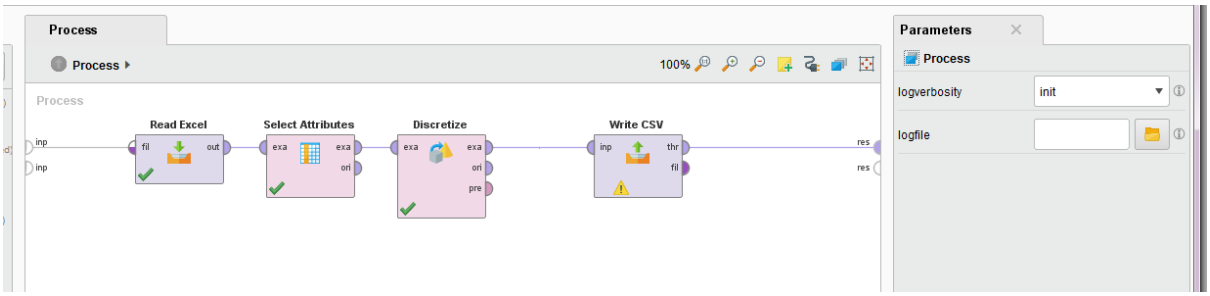
9. หลังจากกด apply แล้ว กด play ให้โปรแกรมทำงาน หลังจากทำงานแล้วเกรดแต่ละตัวที่เป็นตัวเลขจะโดนแปลงเป็นข้อความยกเว้นช่อง endgrade เพราะเรากำหนดเป็น label เราต้องแก้ไขทีหลัง

Row No.	endgrade	ไทย	คณิต	วิทย์	สังคม	สุขศึกษา	ศิลปะ	การงาน	ต่างประเทศ	รหัสนักศึกษา	แผนก
1	2.210	ดีมาก	ปานกลาง	ดีมาก	ดีมาก	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5822041025	คอมพิวเตอร์ธุ...
2	2.490	ปานกลาง	ต่ำ	ต่ำ	ปานกลาง	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	5122010235	การขาย
3	2.280	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010357	การขาย
4	2.500	ต่ำ	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	5122010225	การขาย
5	2.820	ดีมาก	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010315	การขาย
6	2.400	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010336	การขาย
7	2.240	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5022010174	การขาย
8	2.310	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010321	การขาย
9	2.530	ดีมาก	ปานกลาง	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010353	การขาย
10	2.380	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5022010208	การขาย
11	2.370	ต่ำ	ต่ำ	ปานกลาง	ปานกลาง	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010278	การขาย
12	2.560	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010343	การขาย
13	2.970	ปานกลาง	ปานกลาง	ปานกลาง	ดีมาก	ดีมาก	ดีมาก	ต่ำ	ดีมาก	4922010304	การขาย
14	2.640	ปานกลาง	ปานกลาง	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5022010173	การขาย
15	2.630	ดีมาก	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010318	การขาย
16	2.500	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010390	การขาย
17	2.790	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010295	การขาย

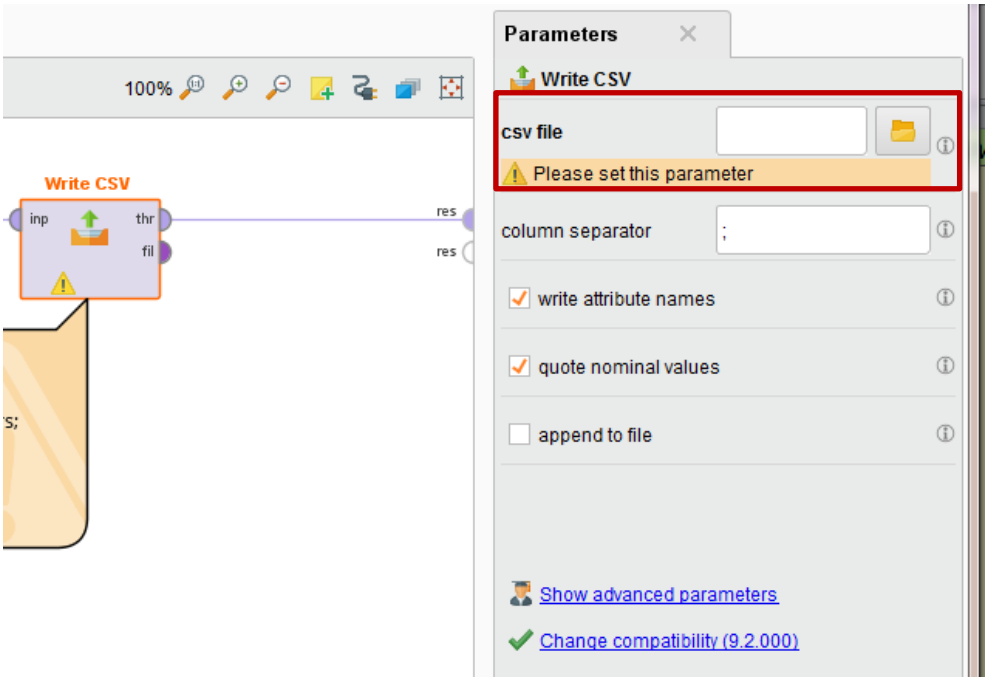
10. หลังจากนั้นเซฟไฟล์ เป็น CSV. เพื่อที่จะสามารถนำมาใช้แล้วนำมาปรับแก้ได้ โดยพิมพ์ค้นหาที่ operator ว่า write Csv.



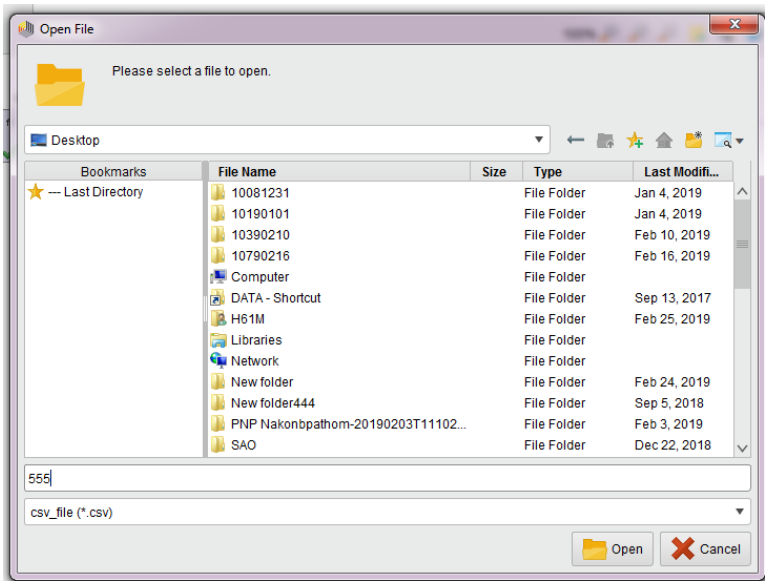
11. นำ Operator write Csv. มาวางที่หน้าจอการทำงาน แล้วลากเส้นเชื่อมกัน ตามรูป



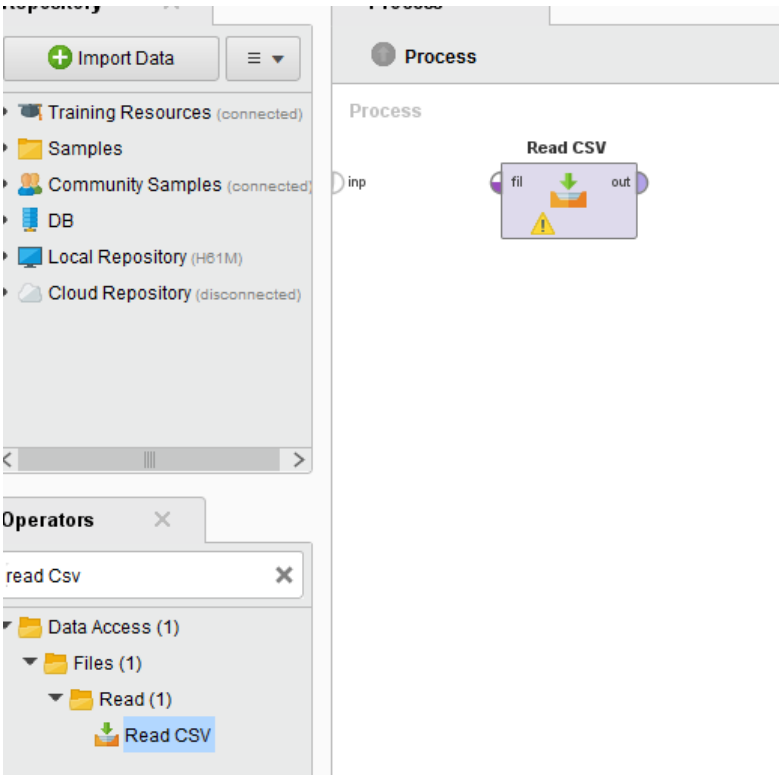
12. หลังจากลากเส้นแล้วเลือก Operator write Csv. แล้วเลือกที่ Save ไฟล์



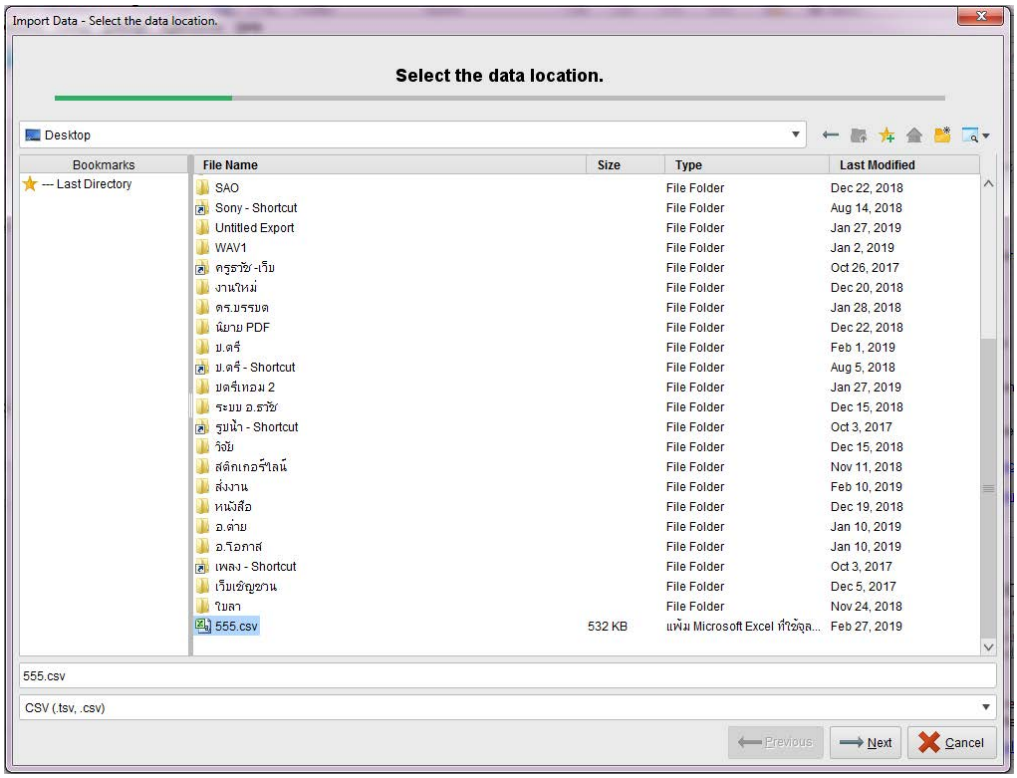
13. เมื่อกดเข้ามาที่ Csv file แล้วจะขึ้นหน้าต่างให้เราเลือก Save file ลงที่ตำแหน่งที่เราต้องการ หลังจากนั้นตั้งชื่อไฟล์และจากนั้นกด Open เมื่อเสร็จแล้วกด Run โปรแกรมให้โปรแกรมทำงาน ไฟล์ Csv. ก็จะถูกบันทึก



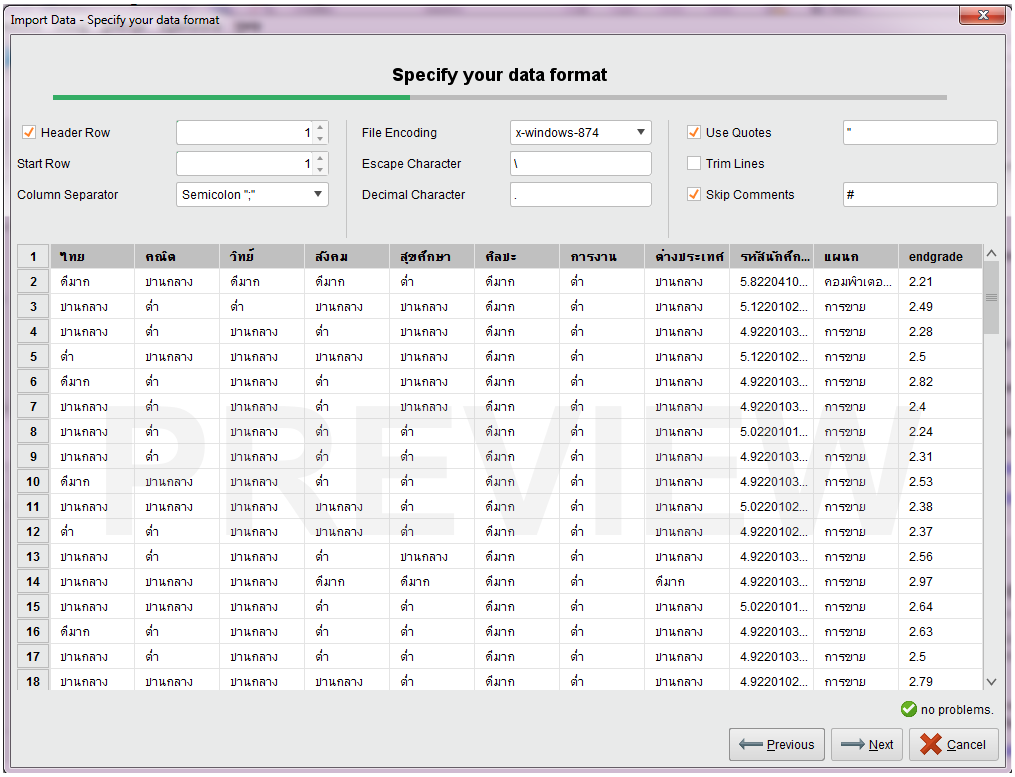
14. หลังจากนั้น เปิด file ขึ้นมาใหม่ ในช่อง Operater พิมพ์ read Csv. เพื่อดึงไฟล์ Csv. ขึ้นมาใช้งาน ดับเบิลคลิกที่ตัว Read Csv.



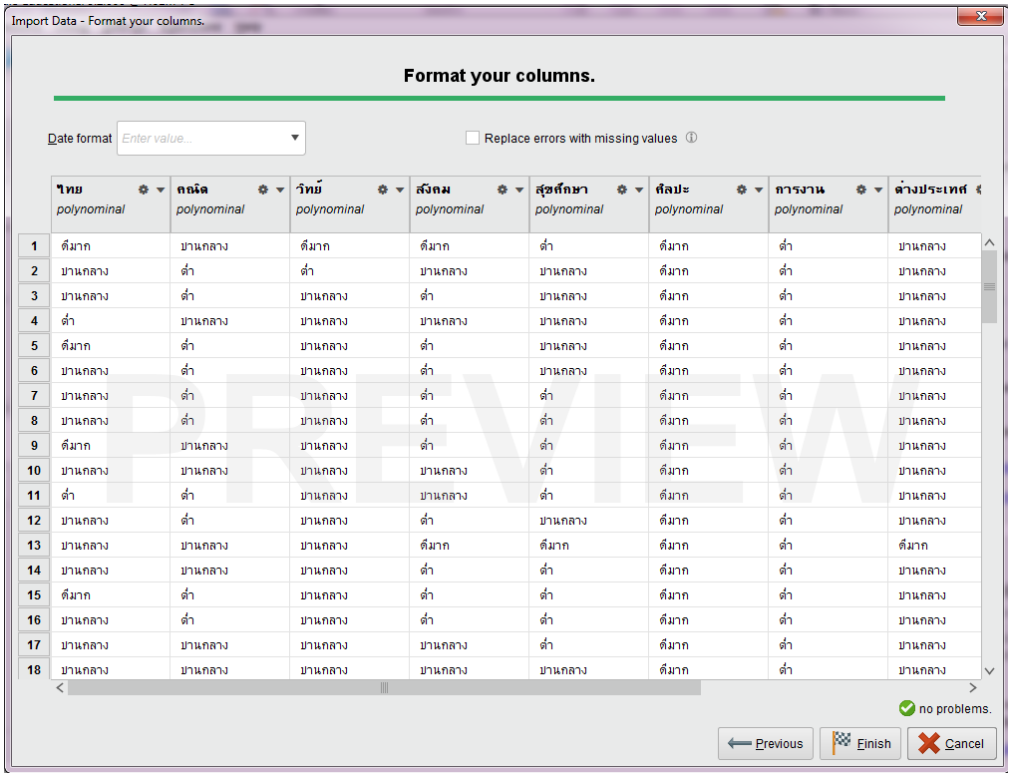
15. โปรแกรมจะแสดงหน้าต่างให้เราเลือกไฟล์ Csv. ที่เราบันทึไว้ก่อนหน้านี้เพื่อนำมาใช้งาน



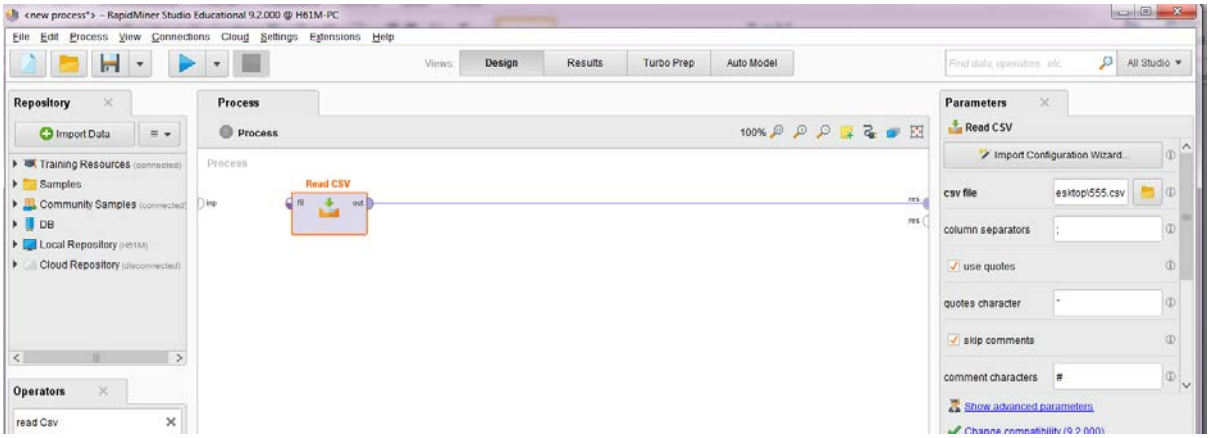
16. หน้าต่างนี้จะแสดง Attributes ที่เราสามารถเลือกที่จะนำมาใช้ได้



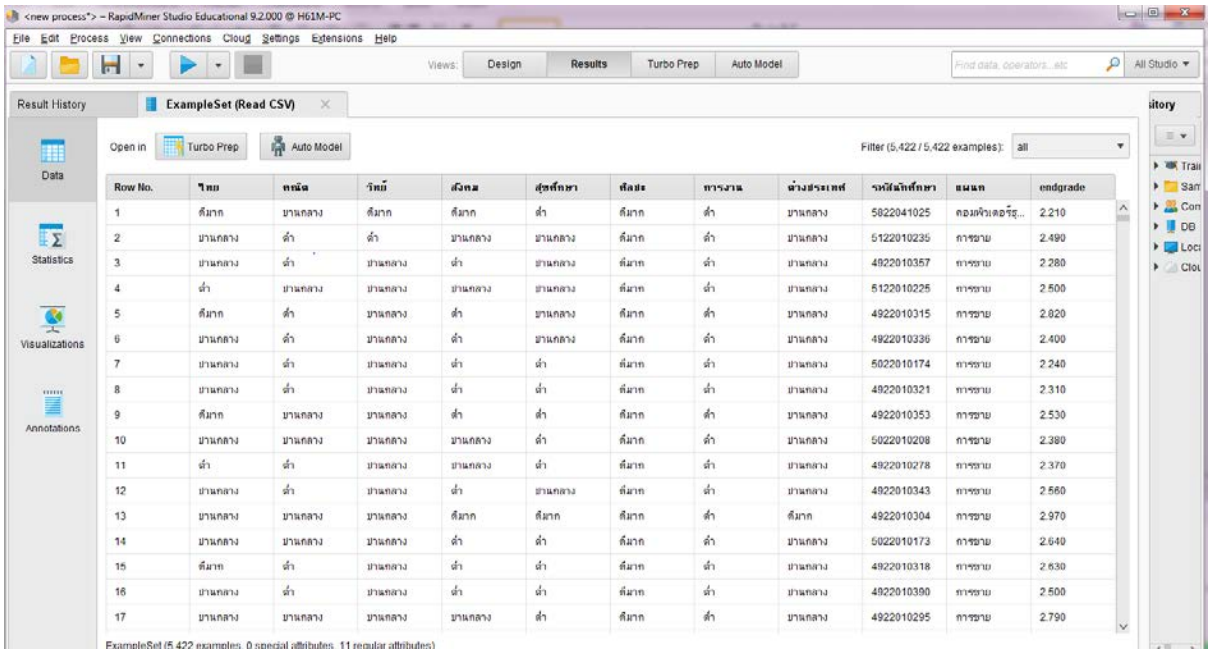
17. หน้าต่างนี้จะเป็นหน้าต่างคุณสมบัติของ Attributes แต่เรายังไม่ต้องกำหนดเพราะเราต้องการเปลี่ยนค่าของ endgrade ให้เป็นไปตามเกณฑ์ที่เรากำหนดไว้ หลังจากนั้นกด finish



18. เมื่อเข้าสู้หน้าต่าง ลากเส้นเชื่อมเพียงเส้นเดียวดังรูป แล้วโปรแกรมลองกด Run Program หน้าต่างจะแสดงรายการเช่นเดียวกับไฟล์ excel ที่เราบันทึกมาก่อนน้ำน้านี้



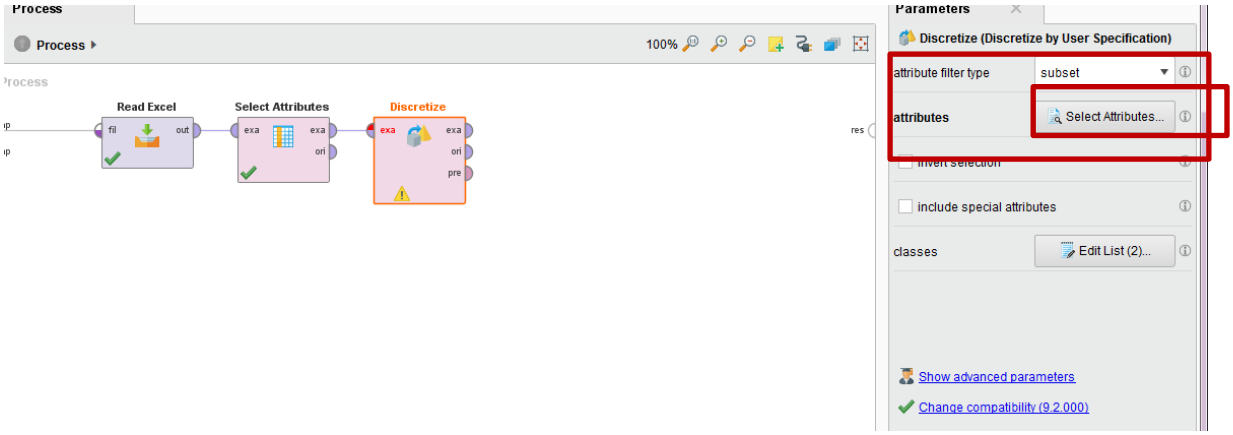
19. หน้าต่างโปรแกรม หลังจากกด Run โปรแกรม



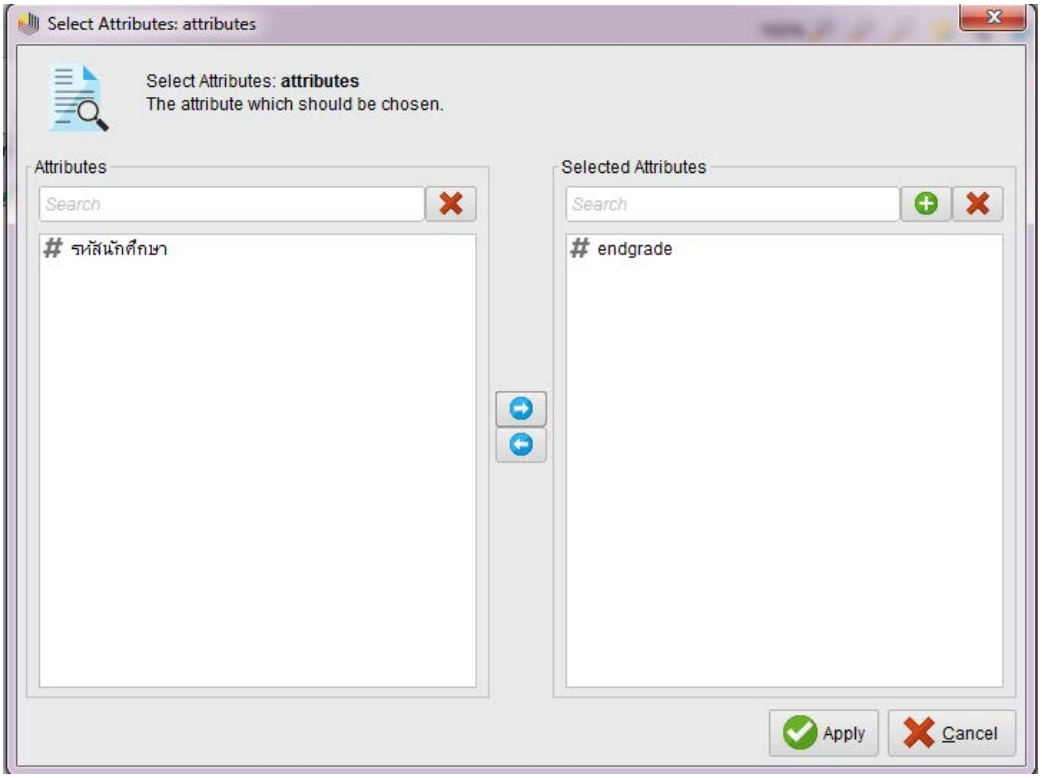
20.หลังจากได้ข้อมูลเบื้อต้นแล้วกลับมาที่หน้าต่าง design เพื่อทำให้ endgrade เปลี่ยนค่าของตัวเลขเป็นข้อความแบ่งระดับ โดยจะแบ่งระดับ เป็น

- ดีมาก = 4
- ปานกลาง = 3
- ต่ำ =2

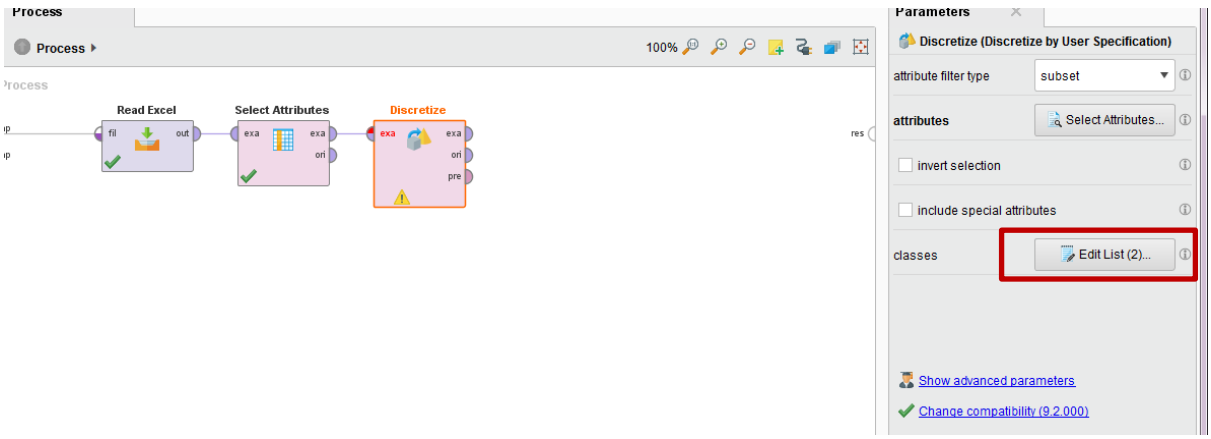
การตั้งค่าเพื่อให้ง่ายต่อการใช้งานโดยการ คลิกขวาเลือก Insert Operator >> Cleansing >> Binning >> Discretize by User Specification



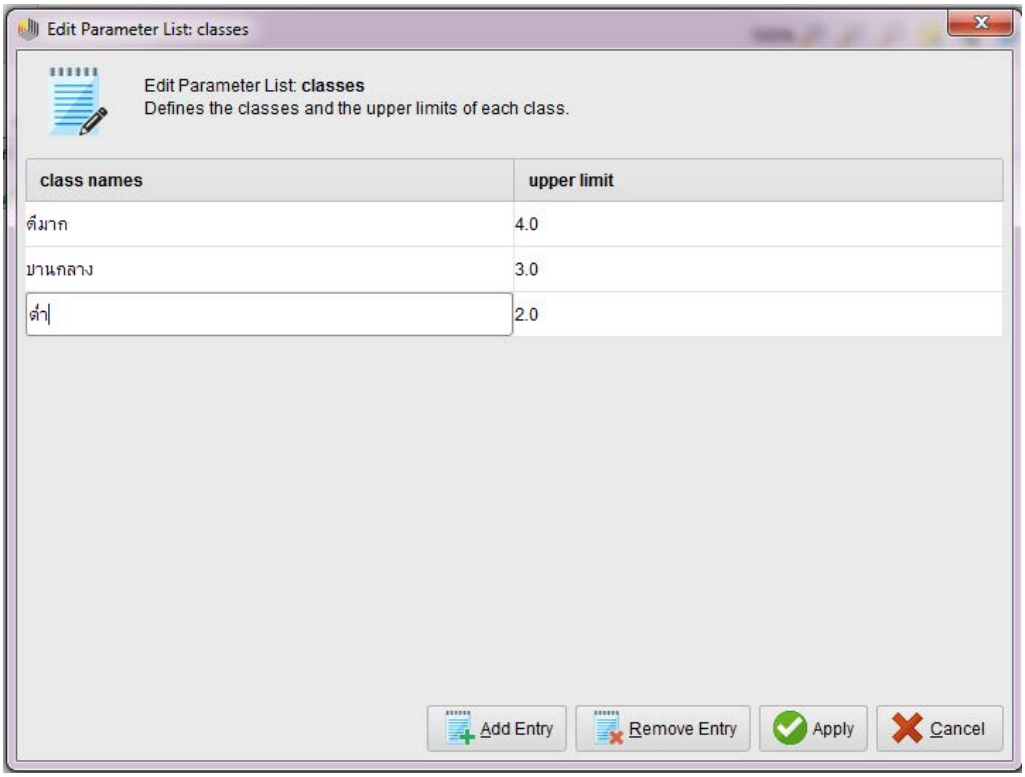
21. เมื่อได้ Operator แล้วเลือกตัว Operator แล้วเลือก Attributes filter type เป็น subset หลังจากนั้นเลือก Select Attributes เพื่อเลือก Attributes ที่ต้องการแทนที่ค่า เมื่อเลือก Attributes ที่ต้องการเสร็จแล้วกด apply



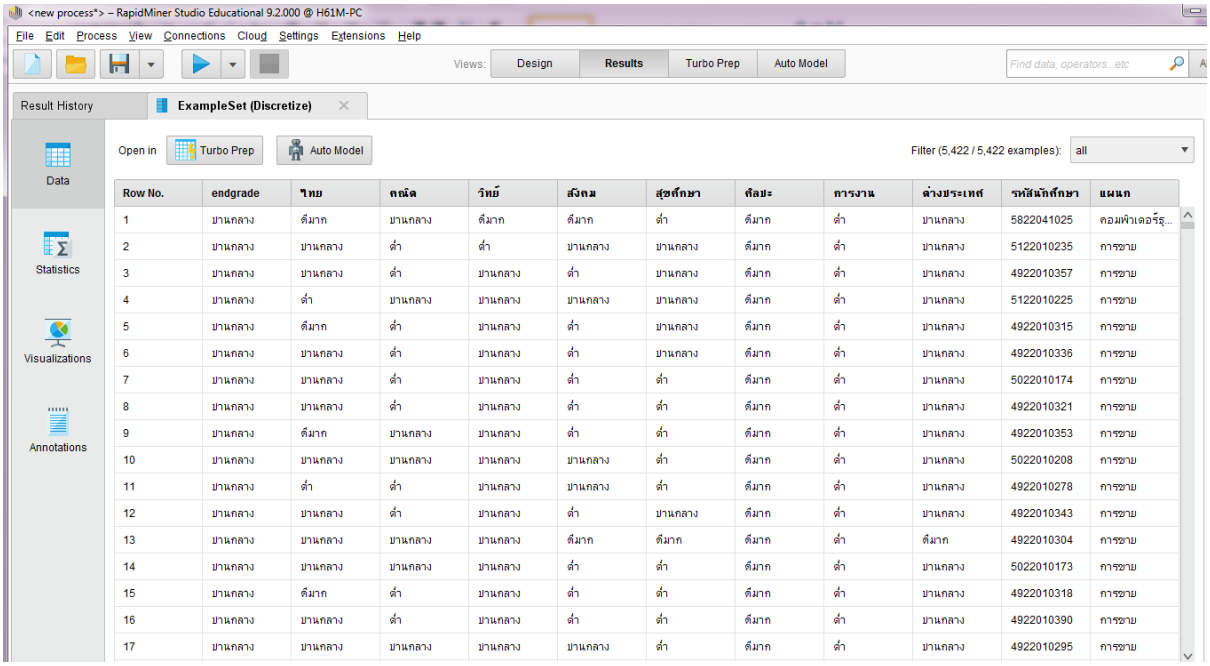
22. หลังจากนั้นกำหนดเกณฑ์ที่ต้องการเทียบกับคะแนน เพื่อแปลงเป็นระดับที่เราตั้งไว้ ในช่อง Classes เลือก Edit List



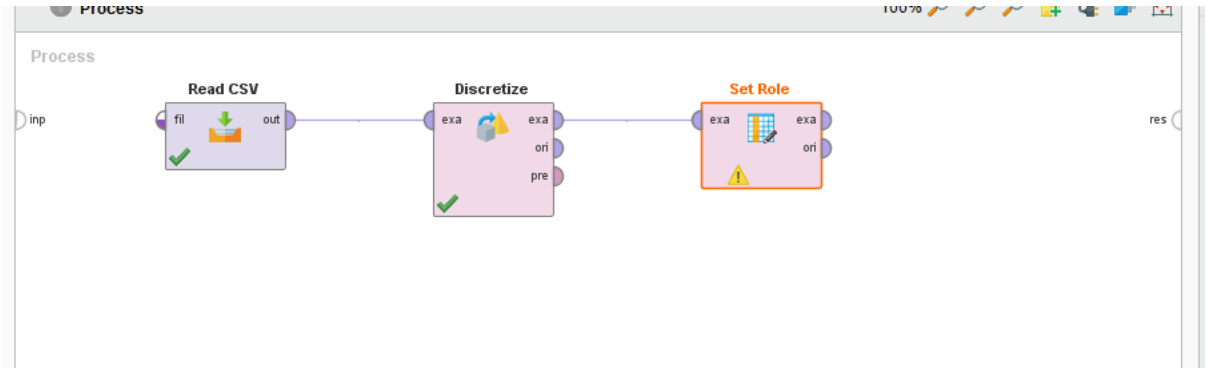
23. เมื่อกดเข้าไปจะมีหน้าต่าง ให้เรากำหนดเกณฑ์ เมื่อเรากำหนดเกณฑ์แล้วกด Apply



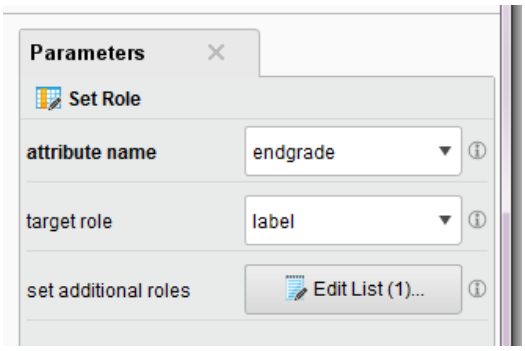
24. หลังจากตั้งค่าเสร็จ กด Apply แล้วสั่งโปรแกรม Run จะได้ผลตามภาพ endgrade เปลี่ยนเป็นข้อความ

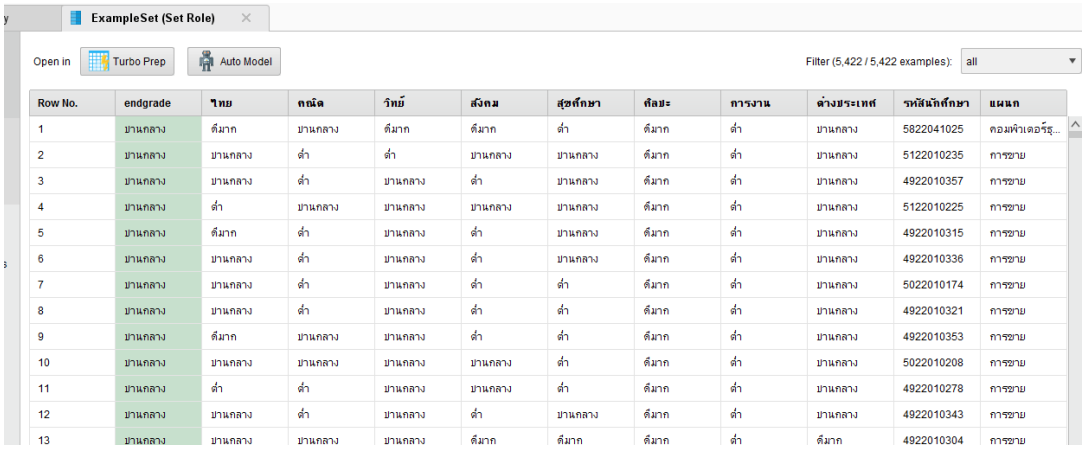


25. หลังจากนั้นเราจะต้องตั้งค่า endgrade ให้เป็น label เพื่อใช้ในการทำนาย โดยพิมพ์ค้นหาที่ operator ว่า Set Role และลาก Operator Set Row มาว่างที่หน้าต่างการทำงาน และลากเส้นให้เชื่อมต่อกัน ดังรูป



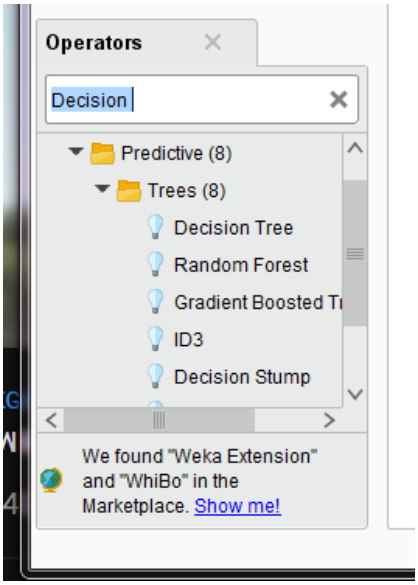
26. หลังจากนั้น คลิกที่ operator Set Role แล้วเลือก attribute name เป็น endgrade เลือก target role เป็น label หลังจากนั้นลากเส้นเชื่อมเส้นสุดท้าย และ กด Run โปรแกรม ตัว endgrade ก็จะเป็นเณฑ์เดียวกับ Attributes อื่น ๆ



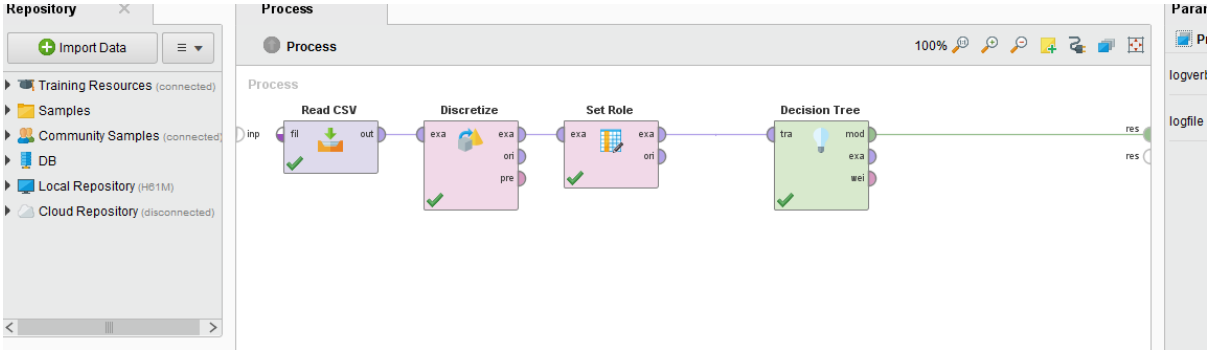


Row No.	endgrade	ไทย	คณิต	วิทย์	สังคม	สุขศึกษา	ศิลปะ	การงาน	ต่างประเทศ	รหัสนักศึกษา	แผนก
1	ปานกลาง	ดีมาก	ปานกลาง	ดีมาก	ดีมาก	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5822041025	ทองพิเดอร์ธุ...
2	ปานกลาง	ปานกลาง	ต่ำ	ต่ำ	ปานกลาง	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	5122010235	การขาย
3	ปานกลาง	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010357	การขาย
4	ปานกลาง	ต่ำ	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	5122010225	การขาย
5	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010315	การขาย
6	ปานกลาง	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010336	การขาย
7	ปานกลาง	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5022010174	การขาย
8	ปานกลาง	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010321	การขาย
9	ปานกลาง	ดีมาก	ปานกลาง	ปานกลาง	ต่ำ	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010353	การขาย
10	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	5022010208	การขาย
11	ปานกลาง	ต่ำ	ต่ำ	ปานกลาง	ปานกลาง	ต่ำ	ดีมาก	ต่ำ	ปานกลาง	4922010278	การขาย
12	ปานกลาง	ปานกลาง	ต่ำ	ปานกลาง	ต่ำ	ปานกลาง	ดีมาก	ต่ำ	ปานกลาง	4922010343	การขาย
13	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ดีมาก	ดีมาก	ดีมาก	ต่ำ	ดีมาก	4922010304	การขาย

27.หลังจากนั้น เลือกมุมมอง Design ต่อไปเราจะทำการสร้างโมเดล Decision Tree โดยการเลือกโอเปอเรเตอร์ Decision Tree จากส่วนของ Operators โดยการพิมพ์ตรงช่องค้นหา โดยพิมพ์คำว่า Decision กดปุ่ม Enter ก็จะปรากฏโอเปอเรเตอร์ Decision Tree ขึ้นมา หรือจะทำการเลือกจากหมวด Modeling >> Classification and Regression >> Tree Induction



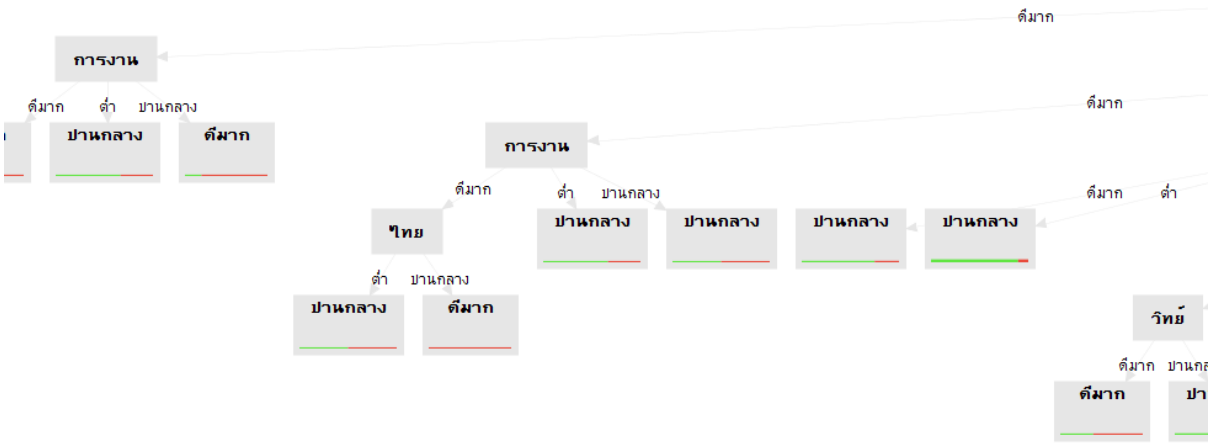
11. ลากโอเปอเรเตอร์ Decision Tree มาวางในส่วนของ Process ตรงเส้นที่เชื่อมต่อเดิมที่โอเปอเรเตอร์ Read Excel ลากไว้ (โปรแกรมจะทำการเชื่อมโอเปอเรเตอร์ทั้งสองตัวทันทีจากพอร์ต out ของโอเปอเรเตอร์ Read Excel ไปยังพอร์ต tra (training) ของโอเปอเรเตอร์ Decision Tree เพื่อเป็นการส่งข้อมูลไปสร้างโมเดล



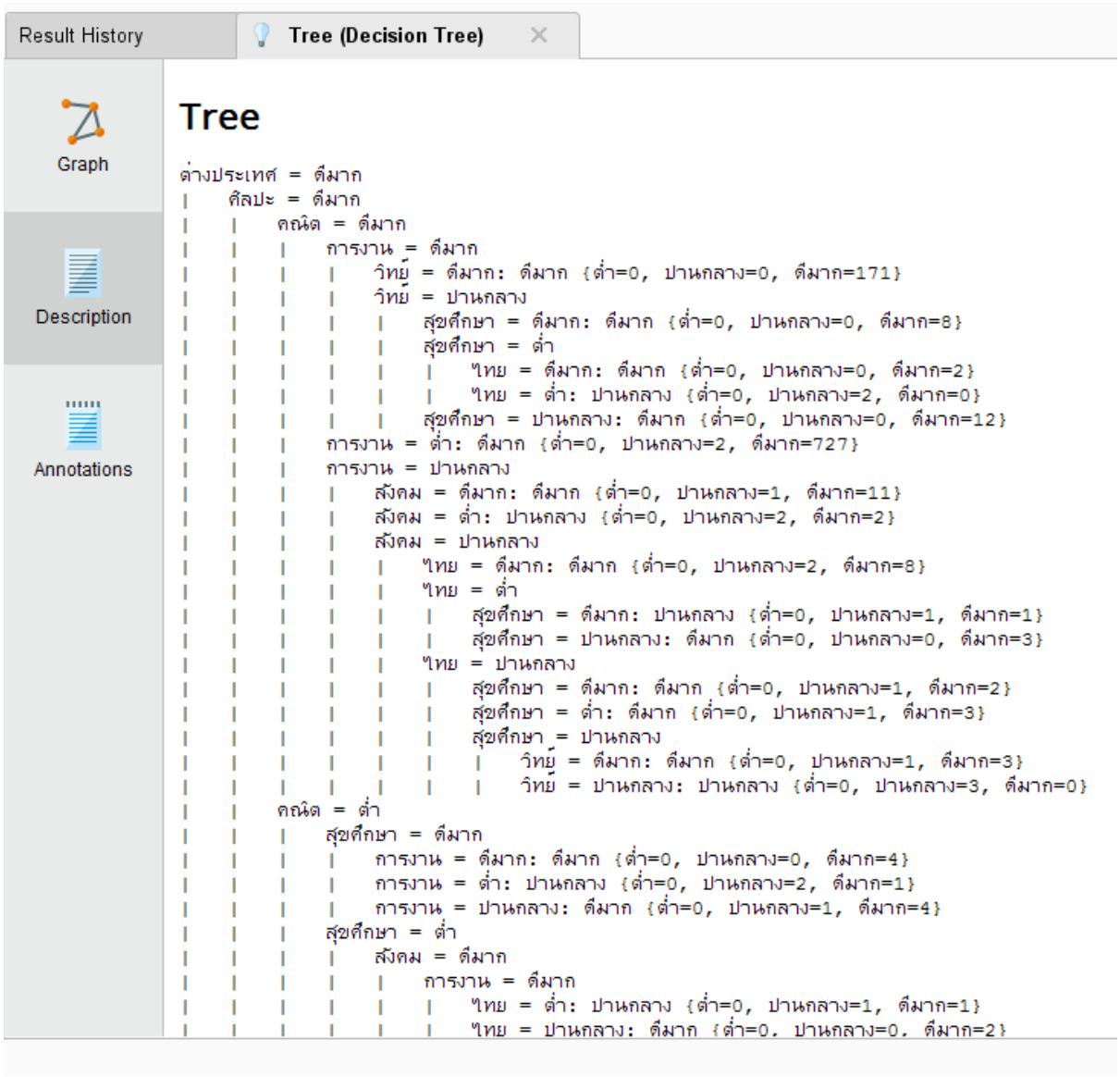
12. ลากเส้นเชื่อมจากพอร์ต mod (model) และพอร์ต exa (example) ของโอเปอเรเตอร์ Decision Tree ไปยังพอร์ต res (result) ทั้งสองพอร์ต เพื่อไปแสดงในส่วนของหน้าจอผลลัพธ์โดยพอร์ต mod จะทำการส่งโมเดล Decision Tree ที่สร้างออกไปแสดงในรูปต้นไม้ และพอร์ต exa จะส่งข้อมูลที่ import เข้ามาไปแสดงในรูปแบบตาราง

13. จากนั้นคลิก Run Process จะได้รูปโมเดลต้นไม้ ซึ่งโมเดลต้นไม้ที่สร้างได้มีส่วนประกอบสำคัญ 3 ส่วน คือ

- ในโมเดล Decision Tree จะมีโหนดต่าง ๆ 2 ประเภท คือ
    - โหนดที่เป็นแอตทริบิวต์แสดงด้วยรูปสี่เหลี่ยมที่มีมุมโค้ง
    - โหนดลาเบลแสดงด้วยรูปสี่เหลี่ยมที่มีกราฟแสดงสีต่าง ๆ อยู่ด้วย ในตัวอย่างนี้มี 2 ลาเบล คือ ดีมาก และ ปานกลาง
- ส่วนของ Zoom ใช้สำหรับย่อขยายรูปโมเดล
- ส่วนของ Mode จะใช้สำหรับปรับโหมดของการใช้งานเมาส์

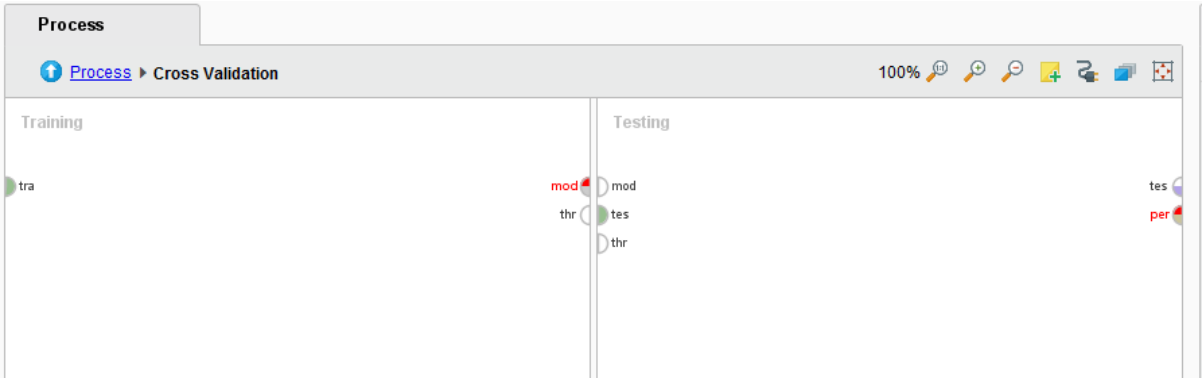


14. ในหน้าต่าง Description จะเป็นโค๊ดข้อความที่เราสามารถนำมาเขียนโปรแกรมเพื่อใช้ในการทำนายได้

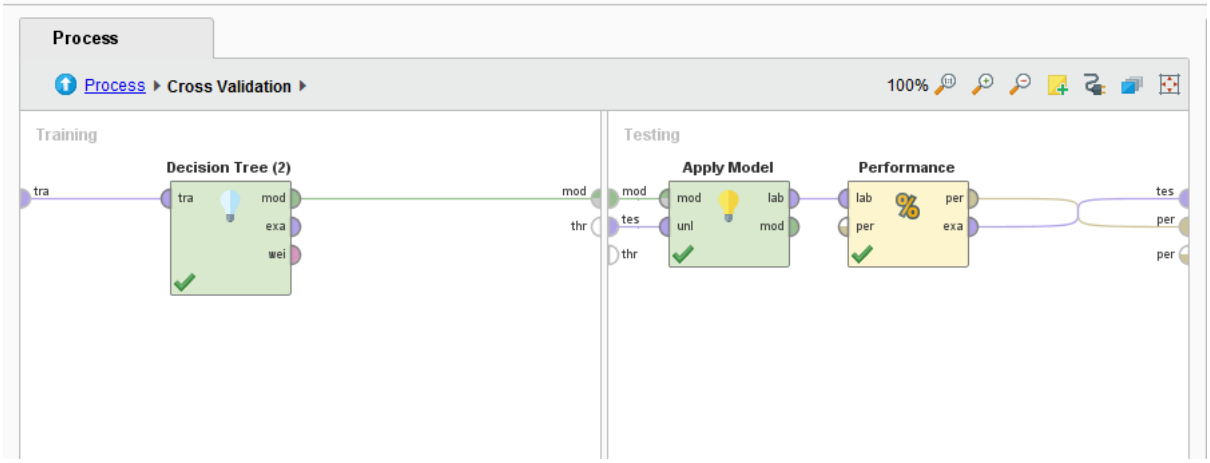


การทดสอบการทำนาย

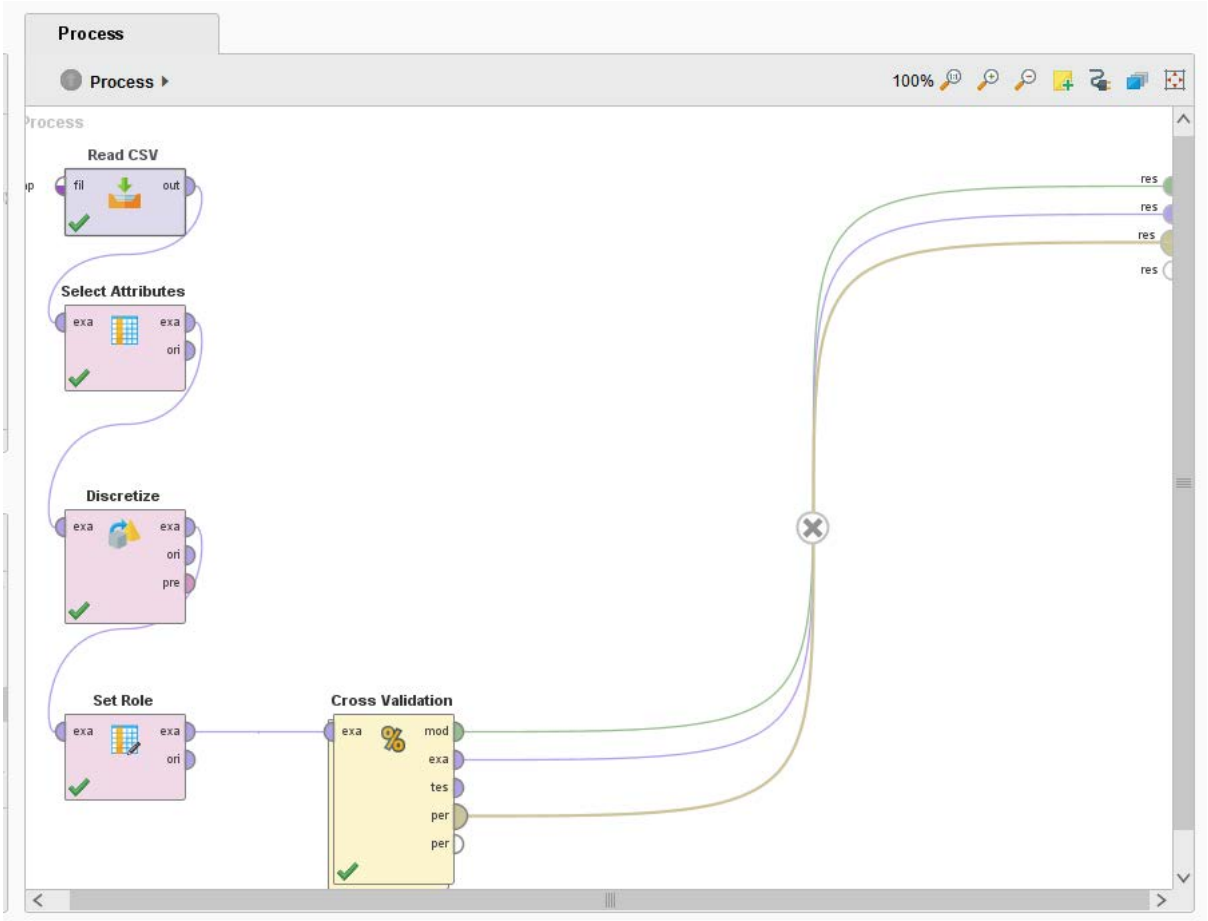
1.การทดสอบการทำนายโดยใช้ Cross Validation ใช้งานโดย คลิกขวาเลือก Insert Operator >> Validation >> Cross Validation หลังจากนั้นดับเบิลคลิกเข้าไปที่ Cross Validation จะแสดงหน้าต่างดังรูป



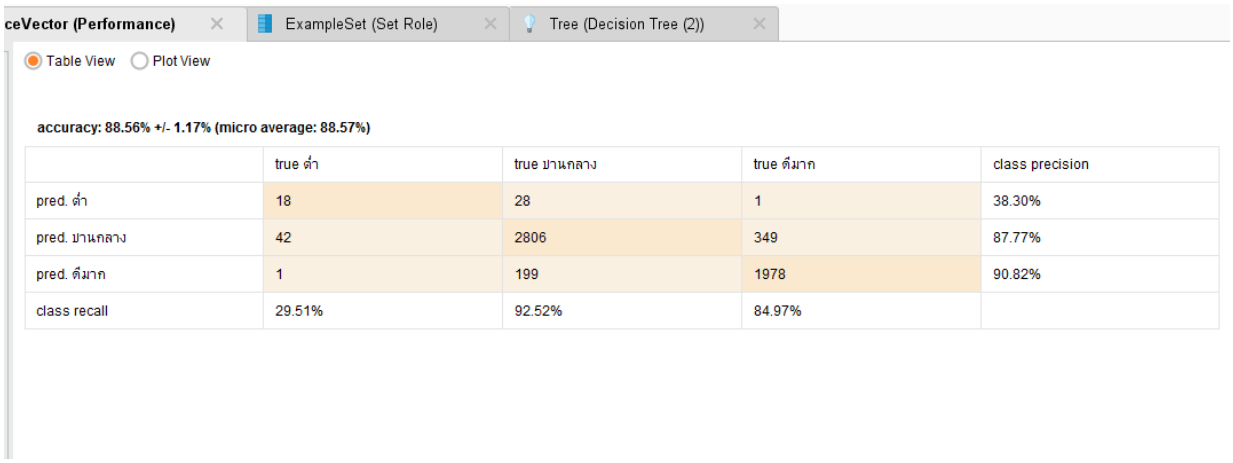
2. ทางด้านซ้ายของช่องให้นำ โมเดล Decision Tree มาวาง และลากเส้นเชื่อม ทางด้านขวา ลาก Apply Model และ Performance มาวาง และลากเส้นดังรูป



3. ที่หน้าต่าง design ก็จะมีการใช้ Operator ต่าง ๆ ดังนี้



4.ผลกาตรวจสอบคุณภาพของการทำนาย หาก% ผลการทำนายยิ่งมากความถูกต้องก็ยิ่งมากขึ้นตามไปด้วย

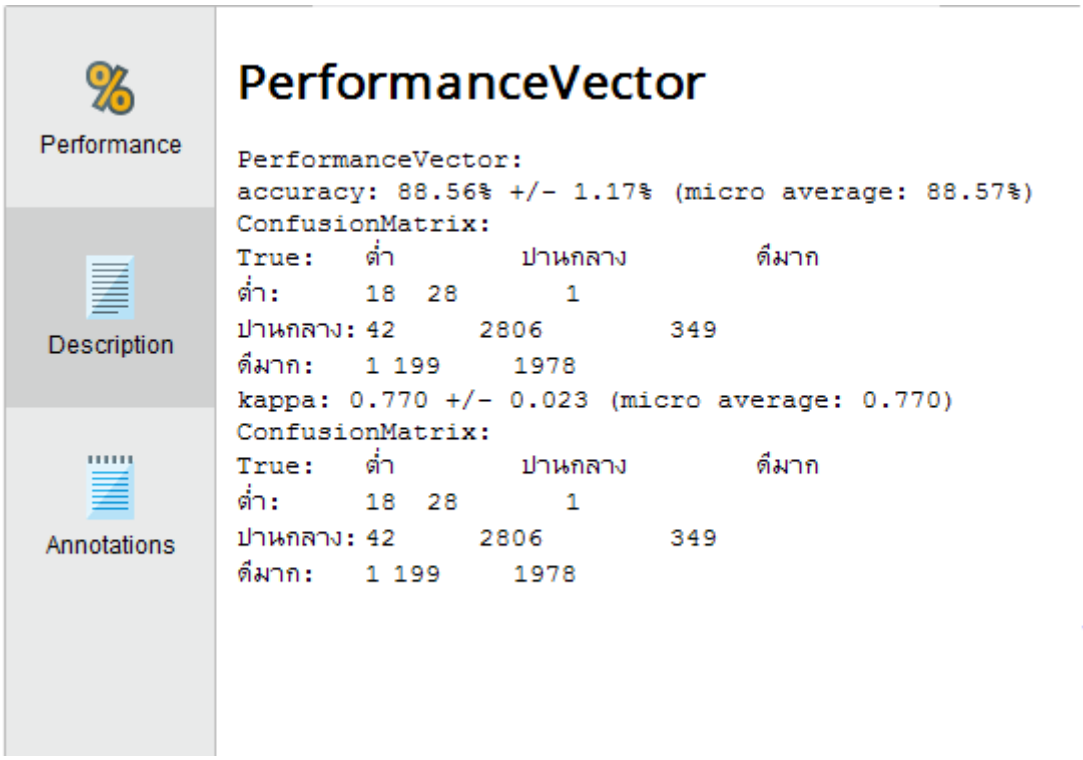


ceVector (Performance) | ExampleSet (Set Role) | Tree (Decision Tree (2))

Table View | Plot View

accuracy: 88.56% +/- 1.17% (micro average: 88.57%)

	true ต่ำ	true ปานกลาง	true ดีมาก	class precision
pred. ต่ำ	18	28	1	38.30%
pred. ปานกลาง	42	2806	349	87.77%
pred. ดีมาก	1	199	1978	90.82%
class recall	29.51%	92.52%	84.97%	



Performance

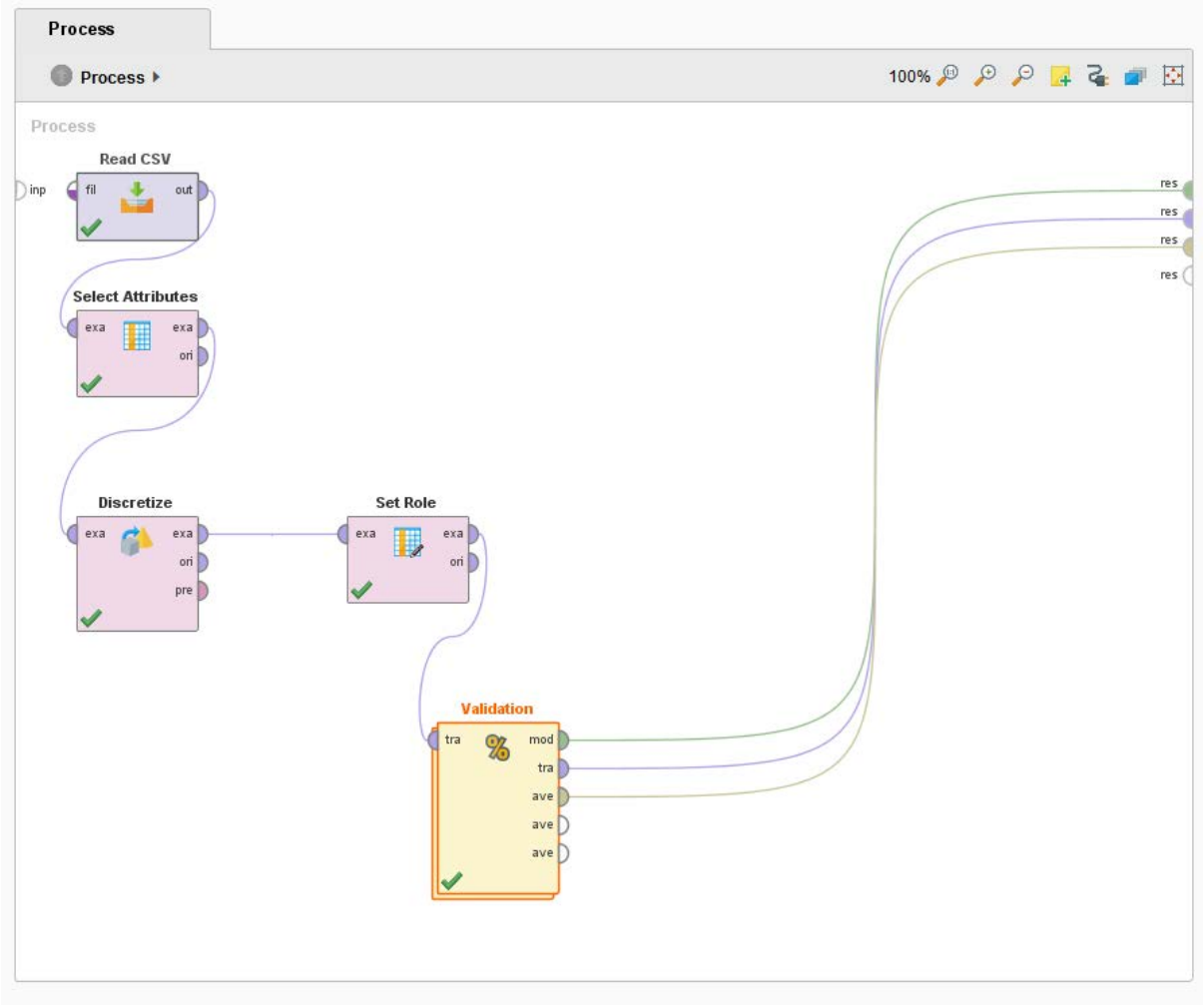
Description

Annotations

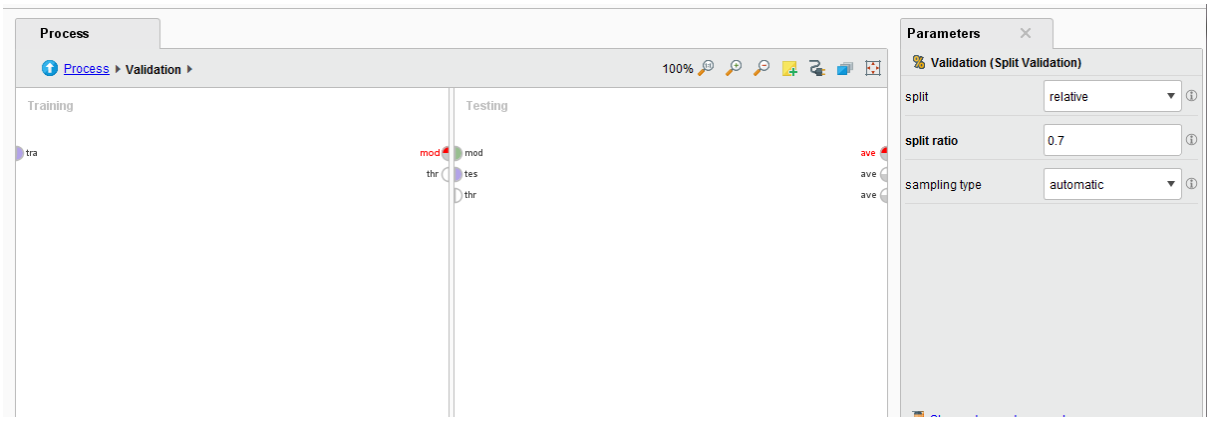
PerformanceVector

```
PerformanceVector:
accuracy: 88.56% +/- 1.17% (micro average: 88.57%)
ConfusionMatrix:
True:    ต่ำ         ปานกลาง         ดีมาก
ต่ำ:      18  28        1
ปานกลาง: 42      2806         349
ดีมาก:    1 199     1978
kappa: 0.770 +/- 0.023 (micro average: 0.770)
ConfusionMatrix:
True:    ต่ำ         ปานกลาง         ดีมาก
ต่ำ:      18  28        1
ปานกลาง: 42      2806         349
ดีมาก:    1 199     1978
```

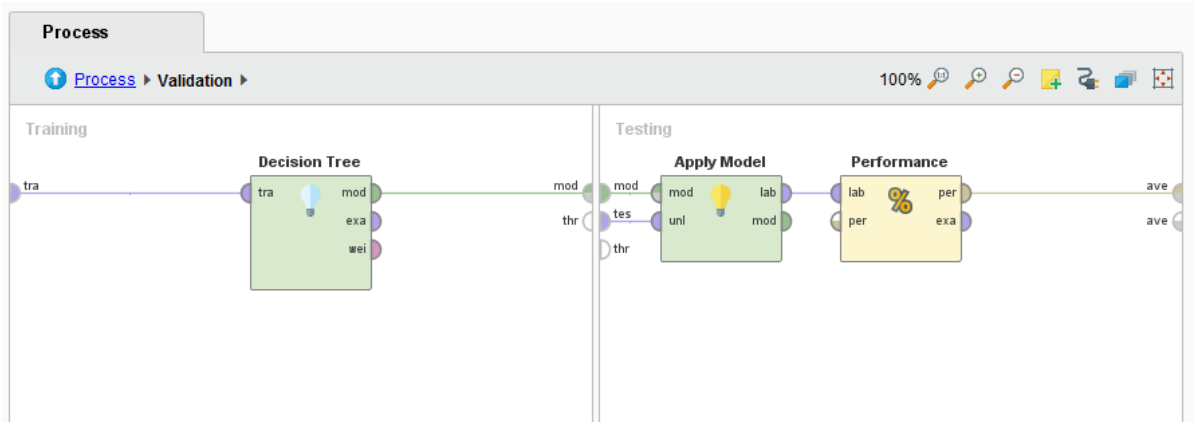
การทดสอบความถูกต้องของผลการทำนายอีกแบบหนึ่งคือการใช้ Split โดยการพิมพ์ค้นหาที่ Operator ว่า Split Validation แล้วเลือกคลิกมาว่างที่หน้าต่างการทำงานแล้วลากเส้นเชื่อมต่อดังภาพ หลังจากนั้นดับเบิลคลิกเข้าไปจะเจอหน้าต่างการทำงาน



หลังจากดับเบิลคลิกเข้ามาที่หน้าต่างแล้วจะเจอหน้าต่างดังภาพ



หลังจากเข้ามาแล้วให้พิมพ์ค้นหาในช่อง Operator และนำ Operator ทางด้านซ้ายของช่องให้นำโมเดล Decision Tree มาวาง และลากเส้นเชื่อม ทางด้านขวา ลากApply Model และ Performance มาวาง และลากเส้นดังรูป



เมื่อกดรันผลการทำนายจะเป็นดังรูป

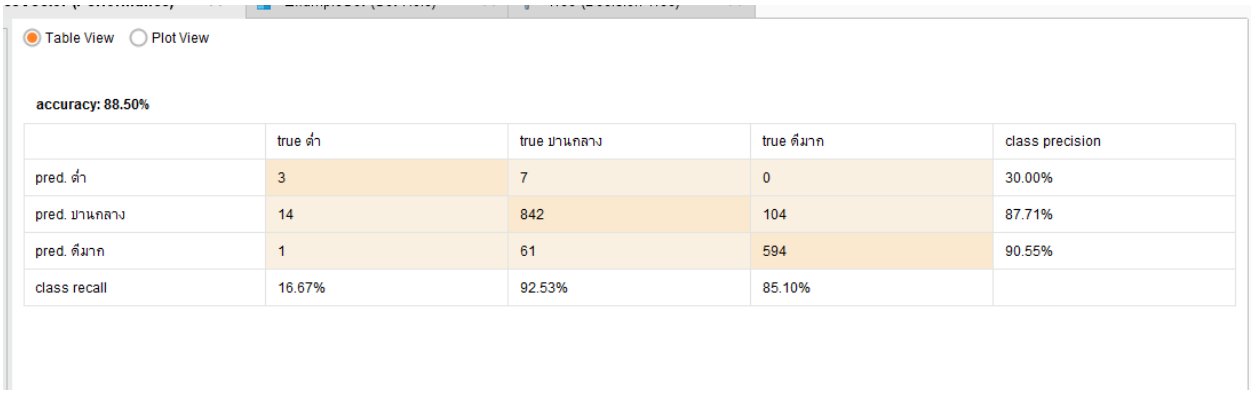
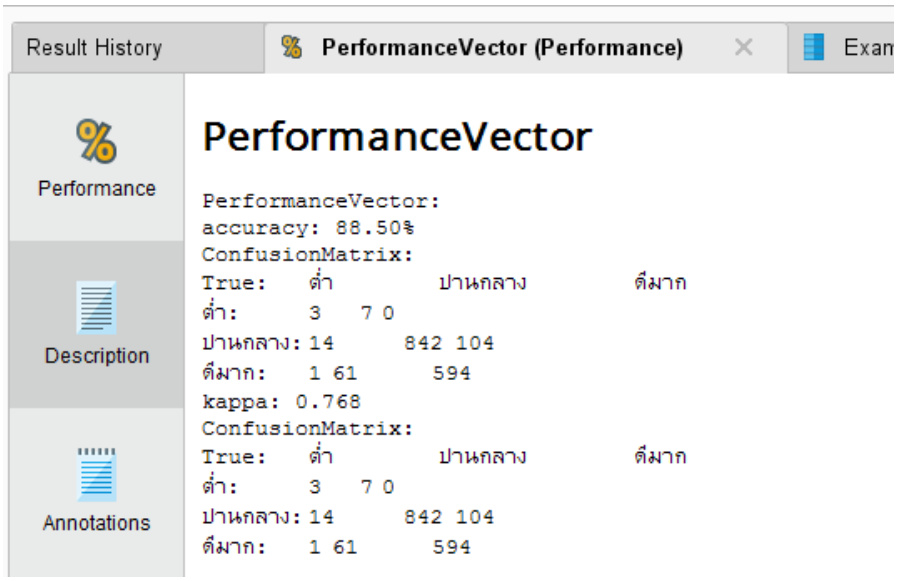


Table View Plot View

accuracy: 88.50%

	true ต่ำ	true ปานกลาง	true ดีมาก	class precision
pred. ต่ำ	3	7	0	30.00%
pred. ปานกลาง	14	842	104	87.71%
pred. ดีมาก	1	61	594	90.55%
class recall	16.67%	92.53%	85.10%	



เอกสารอ้างอิง

http://dataminingtrend.com/2014/wp-content/uploads/2014/02/chapter1.pdf (25-2-62)

http://dataminingtrend.com/2014/wp-content/uploads/2014/02/RM7_chapter1.pdf (25-2-62)

https://behavior.lbl.gov/?q=node/11 (25-2-62)

http://compcenter.bu.ac.th/news-information/data-mining(25-2-62)

rapidminer