การพัฒนาตัวแบบการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิคการทำเหมืองข้อมูล

Development of a Model to Predict Cassava Yield Using Data Mining

ผู้จัดทำ

นายเจ๊ะอานัส ม่องพร้า รหัสนิสิต 592021163

รายงานนี้เป็นส่วนหนึ่งของการศึกษา วิชา เหมืองข้อมูล

หลักสูตร วิทยาศาสตรบัณฑิต วิชาเอกเทคโนโลยีสารสนเทศ

มหาวิทยาลัยทักษิณ ภาคการศึกษา ที่ 2/2562

การพัฒนาตัวแบบการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิคการทำเหมืองข้อมูล*

Development of a Model to Predict Cassava Yield Using Data Mining

เป็นการการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิคการทำเหมืองข้อมูลโดย การจำแนกประเภท (classification) ซึ่งใช้อัลกอริธึมจำนวน 5 ตัว ได้แก่ J48, RandomTree, SimpleCart, NaïveBayes, และLADTree เพื่อสร้างต้นไม้ตัดสินใจ (Decision Tree) จากนั้นทำการทดสอบความแม่นยำใน การพยากรณ์ แล้วคัดเลือกอัลกอริธึมที่ให้ความแม่นยำดีที่สุดไปใช้ในการออกแบบและพัฒนาระบบสารสนเทศ การพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์

วิธีการวิจัย

1. ข้อมูลและกลุ่มตัวอย่างที่ใช้ในการวิจัย

ข้อมูลที่ใช้ในงานวิจัยการพัฒนาตัวแบบพยากรณ์ผลผลิตมันสำปะหลังโดยใช้เทคนิคการทำเหมืองข้อมูล และพัฒนาระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์จะใช้ข้อมูลปัจจัยที่เกี่ยวข้องกับผลผลิตมันสำปะหลังซึ่งเป็นข้อมูลแบบทุติยภูมิตั้งแต่ปี 2551-2555 จากสำนักงานเกษตรจังหวัดกำแพงเพชร ซึ่งประกอบด้วยข้อมูลผลผลิตมันสำปะหลังของจังหวัดในภาคเหนือที่มีการปลูกมันสำปะหลังมากที่สุด และมีสภาพภูมิอากาศที่ใกล้เคียงกับจังหวัดกำแพงเพชร จำนวน 5 จังหวัด ได้แก่ กำแพงเพชร นครสวรรค์ เพชรบูรณ์ พิษณุโลก และอุทัยธานีรวม 325 ครัวเรือน และกลุ่มตัวอย่างในการประเมินความพึงพอใจในการใช้งานระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง ประกอบด้วย เจ้าหน้าที่สำนักงานเกษตรจังหวัดกำแพงเพชร ผู้ใช้งานทั่วไปและผู้ดูแลระบบ รวม 30 คน

2. การทำข้อมูลให้สมบูรณ์

ข้อมูลปัจจัยการผลิตที่นำมาจากสำนักงานเกษตรจังหวัดกำแพงเพชร บางส่วนยังมีความไม่สมบูรณ์จำเป็นต้องดำเนินการทำให้ข้อมูลมีความสมบูรณ์ เพื่อนำมาเข้ากระบวนการในการจัดทำเหมืองข้อมูล โดยทำการจัดข้อมูลให้เป็นรูปแบบมาตรฐานเดียวกัน (Normalization) ปรับเปลี่ยนค่าของแอตทริบิวต์ให้เป็นหน่วยนับเดียวกัน เช่น หน่วย ไร่ งาน ตารางวา ให้เป็นหน่วย ไร่ ปริมาณปุ๋ย ปรับหน่วยเป็นกิโลกรัม ส่วนปริมาณผลผลิตปรับหน่วยเป็นตัน และผลผลิตให้อยู่ในหน่วยมาตรฐาน คือ ตันต่อไร่ หลังจากนั้นกำจัดข้อมูลที่ในแถวเป็นค่าว่าง (NULL) มีค่าข้อมูลเป็น 0 หรือระเบียนที่มีค่าผลผลิตเป็น 0 ออกและเลือกแอตทริบิวต์ที่มีคาดว่าสามารถนำมาใช้ประโยชน์ได้ และมีข้อมูลครบถ้วน เช่น จังหวัด อำเภอ ตำบล พันธุ์ที่ใช้ อายุปลูกถึงเก็บ เดือนเก็บผลผลิต ปีเก็บผลผลิตเนื้อที่เก็บเกี่ยว เนื้อที่เสียหาย เนื้อที่ปลูกเนื้อที่ ชนิดปุ๋ย ปริมาณปุ๋ย และปริมาณผลผลิต เป็นต้น จากนั้นทำการกำหนดรหัสข้อมูล

ชื่อข้อมูล	ความหมาย
Year	ปี
Province	จังหวัด
Amphoe	อำเภอ
Tambon	ตำบล
Village	หมู่บ้าน
Species	พันธุ์ที่ใช้
CassAge	อายุปลูกถึงเก็บ (เดือน)
AreaGrownV	เนื้อที่ปลูก (ตร.ว.)
AreaGrownR	เนื้อที่ปลูก (ไร่)
AreaHarvestV	เนื้อที่เก็บเกี่ยว (ตร.ว.)
AreaHarvestR	เนื้อที่เก็บเกี่ยว (ไร่)
YieldProdTN	ปริมาณผลผลิต (ตัน)
YieldProdKG	ปริมาณผลผลิต (กิโลกรัม)
YieldProdAndAreaGrownx400TN	ผลผลิตต่อไร่ (ตัน/ไร่)
YieldProdAndAreaGrownxNEW	ระยะช่วงผลผลิต (ตัวอักษร)
YieldProdAndAreaGrownxNUM	ระยะช่วงผลผลิต (ตัวเลข)
AreaDamageV	เนื้อที่เสียหาย (ตร.ว.)
AreaDamageR	เนื้อที่เสียหาย (ไร่) (งาน)
FertAreaCHV	เนื้อที่ใส่ปุ๋ยเคมี (ไร่)
FertAreaCHR	เนื้อที่ใส่ปุ๋ยเคมี (งาน)
FertQuanCHKG	ปริมาณปุ๋ยเคมีที่ใช้ (กก.)
FertQuanCHx400	ปริมาณปุ๋ยเคมีต่อเนื้อที่ใช้ปุ๋ยเคมี (ไร่)
FertAreaCKV	เนื้อที่ใส่ปุ๋ยคอก (ตร.ว.)
FertAreaCKR	เนื้อที่ใส่ปุ๋ยคอก (ไร่)
FertQuanCK(KG)	ปริมาณปุ๋ยคอกที่ใช้ (กก.)
FertQuanCLx400	ปริมาณปุ๋ยอินทรีย์ต่อเนื้อที่ปุ๋ยอินทรีย์ (ไร่)
FertQuanCKx400	ปริมาณปุ๋ยคอกต่อเนื้อที่ใช้ปุ๋ยคอก (ไร่)

3. การคัดเลือกข้อมูล

การคัดเลือกข้อมูลปัจจัยการผลิตมันสำปะหลัง ได้พิจารณาจาก การสอบถามข้อมูลเจ้าหน้าที่สำนักงานเกษตรจังหวัดกำแพงเพชร และจากเกษตรกรผู้ปลูกมันสำปะหลัง จึงได้คัดเลือกข้อมูล

แอตทริบิวต์ที่ เกี่ยวข้องกับมันสำปะหลัง จำนวน 11 ปัจจัย ได้แก่ จังหวัด พันธุ์ที่ใช้ อายุปลูก ผลผลิตต่อไร่ เนื้อที่ปลูก เนื้อที่ เก็บเกี่ยว เนื้อที่เสียหาย ปริมาณปุ๋ยเคมี ปริมาณปุ๋ยคอก ปริมาณปุ๋ยชีวภาพ และปริมาณปุ๋ยอินทรีย์ แล้วทำการวิเคราะห์สัมพันธ์ของข้อมูลทั้ง 11 ปัจจัย ทั้งนี้จะไม่นำจังหวัดมาวิเคราะห์หาความสัมพันธ์เนื่องจากจังหวัดมีลักษณะเป็นการแบ่งเขตพื้นที่การปลูกซึ่งก็ชัดเจนว่าเป็นตัวแปรที่มีความสำคัญและสัมพันธ์กับตัวแปรอื่นๆ อยู่แล้ว จากนั้นจะกำหนดให้ผลผลิตต่อไร่เป็นตัวแปรตาม ส่วนปัจจัยการผลิตที่เหลืออีก 9 ปัจจัย กำหนดให้เป็นตัวแปรต้น ซึ่งผลการเคราะห์สหสัมพันธ์ (Correlation Analysis) โดยใช้โปรแกรม SPSS

ปัจจัยการผลิต	ค่าสัมประสิทธิ์สหสัมพันธ์
พันธุ์ที่ใช้	-0.194*
อายุปลูก	0.228*
เนื้อที่ปลูก	0.047**
เนื้อที่เก็บเกี่ยว	0.051**
เนื้อที่เสียหาย	-0.025
ปริมาณปุ๋ยเคมี	0.084*
ปริมาณปุ๋ยคอก	-0.014
ปริมาณปุ๋ยชีวภาพ	0.000
ปริมาณปุ๋ยอินทรีย์	0.025*

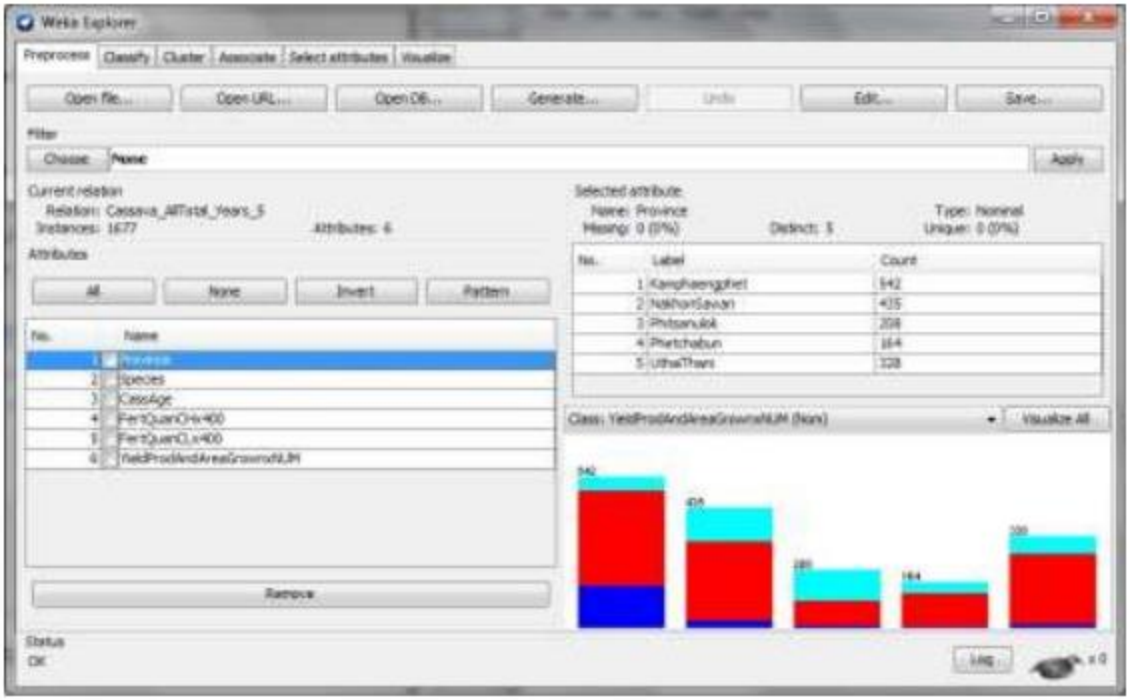
4. การแปลงข้อมูล

จากข้อมูลปัจจัยการผลิตของสำนักงานเกษตรจังหวัดกำแพงเพชร จำนวน 1,764 ระเบียน 40 แอตทริบิวต์ ผู้วิจัยได้ทำการจัดข้อมูลให้มีความสัมบูรณ์และทำการคัดเลือกข้อมูลได้ออกมาเป็นจำนวน 1,677 ระเบียน 11 แอทริบิวต์จากนั้นทำการวิเคราะห์หาความสัมพันธ์ของปัจจัยการผลิตจำนวน 11 ปัจจัย (แอตทริบิวต์) จะได้ปัจจัยการผลิตที่มีความสัมพันธ์กันจำนวน 5 ตัวแปร (แอตทริบิวต์) ทั้งนี้งานวิจัยนี้ได้ใช้วิธีการทำเหมืองข้อมูล โดยใช้เทคนิคต้นไม้ตัดสินใจในการสร้างตัวแบบการพยากรณ์ ซึ่งผลลัพธ์ที่ได้จะแสดงออกมาเป็นค่าระดับ หรือค่าช่วง ฉะนั้น ผู้วิจัยจึงทำการกำหนดค่าระดับผลผลิตผลิตมันสำปะหลัง โดยอาศัยการสอบถามข้อมูลค่าเฉลี่ยผลผลิตสูงสุด และต่ำสุด ต่อพื้นที่การปลูกหนึ่งไร่ จากเจ้าหน้าที่สำนักงานเกษตรจังหวัดกำแพงเพชร และเกษตรกรผู้ปลูกมันสำปะหลัง ซึ่งได้ค่าระดับ ผลผลิตดังนี้

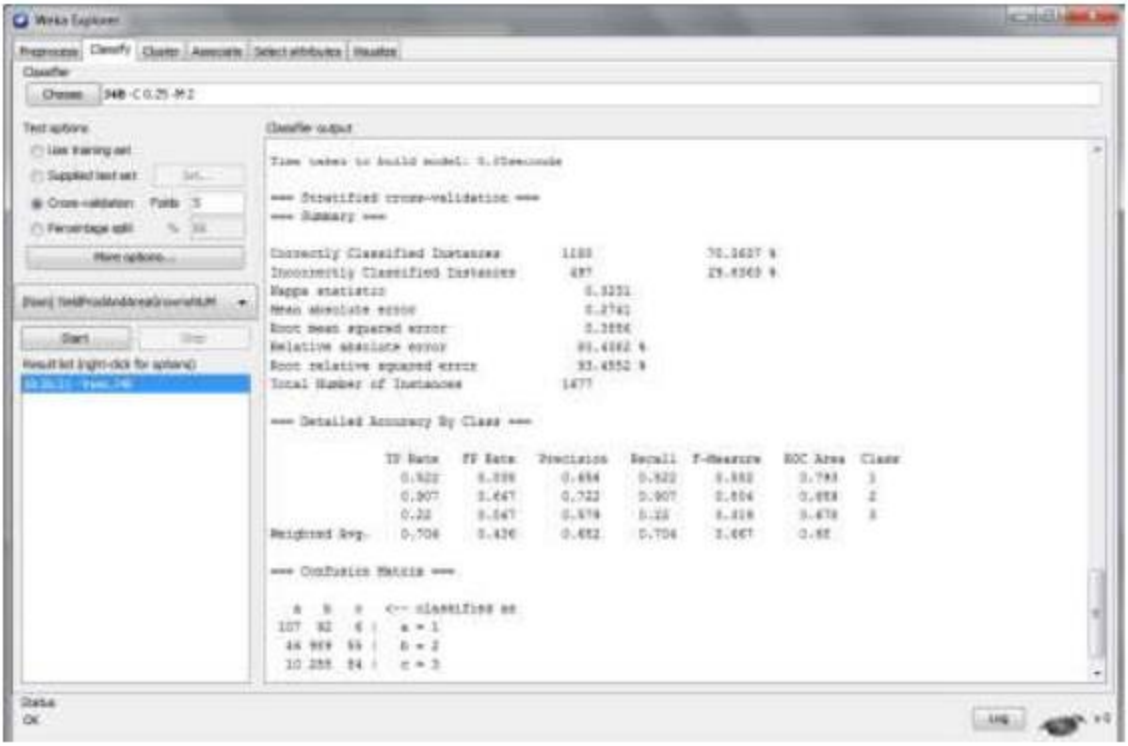
ระดับ 1	= ผลผลิตมากกว่า	6	ตัน/ไร่
ระดับ 2	= ผลผลิตตั้งแต่	3 – 6	ตัน/ไร่
ระดับ 3	= ผลผลิตตั้งแต่	0 – 3	ตัน/ไร่

5. การพัฒนาตัวแบบพยากรณ์

ใช้โปรแกรมเวกา (WEKA) เพื่อสร้างตัวแบบการพยากรณ์ผลผลิตมันสำปะหลัง โดยใช้ข้อมูลปัจจัยการผลิตจำนวน 1,677 ระเบียน 5 แอตทริบิวต์ โดยใช้เทคนิคการจำแหนกประเภทข้อมูล (Classification)ด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) และใช้อัลกอริทึม J48, RandomTree, SimpleCart, NaïveBayes และ LADTree ซึ่งในการนำเข้าข้อมูลสู่โปรแกรมเวกา (WEKA) ต้องทำการแปลงข้อมูลให้อยู่ในรูปของเท็กซ์ไฟล์ (TextFile)



ทำการทดสอบความแม่นยำในการพยากรณ์ด้วยเทคนิค Cross-validation Test โดยทำการแบ่งข้อมูลออกเป็น 5 ส่วน (5-fold cross-validation) และ 10 ส่วน (10-fold cross-validation) จากตัวแบบการพยากรณ์ที่ได้จากการใช้เทคนิคการจำแหนกประเภทข้อมูล (Classification) ด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) ใช้อัลกอริทึม J48, RandomTree, SimpleCart, NaïveBayes และ LADTree



แสดงขั้นตอนการทดสอบความแม่นยำในการพยากรณ์ด้วยเทคนิค 5-fold crossvalidation ของอัลกอริทึม J48 ซึ่งได้ค่าแม่นยำ 70.36% จากนั้นจะทำการทดสอบเช่นนี้ในทุกอัลกอริทึม และเมื่อทำการทดสอบด้วยเทคนิค 5-fold cross-validation แล้วก็จะดำเนินการทดสอบด้วยเทคนิค

10-fold cross-validation จากนั้นจะทำการปรับปรุงวิธีการทดสอบให้มีความแม่นยำที่ดีขึ้น โดยทำการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลเรียนรู้ (Training Set) และข้อมูลทดสอบ (Test Set) (Data Set = Training Set+ Test Set) โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามสัดส่วนดังกล่าว ทำให้จะได้ชุดข้อมูลที่จะทำการทดสอบจำนวน 5 ชุด แล้วทำการทดสอบจากอัลกอริทึม J48, RandomTree, SimpleCart, NaïveBayes และ LADTree

6. การออกแบบระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์

7. สรุปและอภิปรายผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาตัวแบบการพยากรณ์ผลผลิตมันสำปะหลังด้วยเทคนิคการทำเหมืองข้อมูล และเพื่อพัฒนาระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจของผู้ใช้งานระบบผ่านเว็บไซต์ โดยการพัฒนาตัวแบบการพยากรณ์จะใช้ตัวแบบการพยากรณ์ที่ได้จากโปรแกรมเวกา (WEKA) โดยใช้เทคนิคการทำเหมืองข้อมูลจะใช้การจำแนกประเภท(classification) และใช้อัลกอริธึมจำนวน 5 ตัว ได้แก่ J48, RandomTree, SimpleCart, NaïveBayes, และ LADTree เพื่อสร้างต้นไม้ตัดสินใจ (Decision Tree) จากนั้นทำการทดสอบความแม่นยำในการพยากรณ์ด้วย เทคนิค Cross-validation Test โดยทำการแบ่งข้อมูลออกเป็น 5 ส่วน (5-fold cross-validation) และ 10 ส่วน (10-fold cross-validation) ซึ่งผลการทดสอบพบว่าทั้งสองเทคนิคยังให้ค่าความแม่นยำสูงสุดเพียง70.96% ดังนั้นผู้วิจัยจึงทำการปรับปรุงวิธีการทดสอบให้มีความแม่นยำที่ดีขึ้น โดยการทำการแบ่งข้อมูลออกเป็น2 ส่วน ได้แก่ ข้อมูลเรียนรู้ (Training Set) และข้อมูลทดสอบ (Test Set) (Data Set = Training Set + TestSet) จำนวน 5 ชุด ผลทดสอบพบว่ามีความแม่นยำเพิ่มขึ้นสูงสุดถึง 80.12% โดยสรุปได้ดังนี้ อัลกอริทึมของชุดข้อมูล Test Set ที่ได้ความแม่นยำสูงสุดจำนวน 3 อัลลอริทึม ได้แก่ ชุดข้อมูลที่ 5 อัลกอริทึม J48 ให้ค่าความแม่นยำที่ 75.64% ชุดข้อมูลที่ 5 ของอัลกอริทึม SimpleCart ให้ค่าความแม่นยำที่ 80.12% และชุดข้อมูลที่ 4 ของอัลกอริทึม LADTree ให้ค่าความแม่นย าที่ 78.68% ส่วนอัลลอริทึมที่เหลือได้แก่ อัลกอริทึม RandomTree ให้ค่าความแม่นยำที่ 66.85% และอัลกอริทึม NaïveBayes ให้ค่าความแม่นยำที่ 67.91% ซึ่งให้ค่าความแม่นยำที่ต่ำจึงไม่นำมาใช้ในการสร้างตัวแบบการพยากรณ์ จากนั้นจึงดำเนินการนำอัลกอริธึมที่ให้ความแม่นยำดีที่สุดไปใช้ในการออกแบบและพัฒนาระบบ

สารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์ต่อไป

ซึ่งสอดคล้องกับงานวิจัยของปฏิยากร โคผดุง (2551) ได้นำเสนองานวิจัยเรื่องต้นไม้ตัดสินใจสำหรับการวิเคราะห์ภาวะตัวรับฮอร์โมนของผู้ป่วยมะเร็งเต้านมและอัตราการอยู่รอดชีวิตของผู้ป่วยมะเร็งเต้านมในโรงพยาบาลขอนแก่น โดยการจำแนกระดับภาวะตัวรับฮอร์โมนสาหรับแนวทางการรักษาผู้ป่วยมะเร็งเต้านม และใช้เป็นระบบสนับสนุนการตัดสินใจด้านคลินิก โดยการประยุกต์ใช้อัลกอริทึม J48 ในโปรแกรมเวกา สร้างตัวแบบต้นไม้ตัดสินใจโดยใช้ทฤษฎีการทำเหมืองข้อมูลกับข้อมูลผู้ป่วยจำนวน 57 คน ซึ่งทำการวิเคราะห์ข้อมูลจำนวน 57 instance มีแอททริบิวต์จำนวน 6 แอททริบิวต์สำหรับการประเมินความถูกต้องของตัวแบบที่ใช้ทำนายใช้วิธี 10-fold cross validation และวิธีTrain test พบว่าวิธี10-fold cross validation มีความถูกต้อง 57.89 %และ วิธีTrain test มีความถูกต้อง 66.67% ซึ่งวิธีนี้จะมีประสิทธิภาพสูงสุด ซึ่งผู้วิจัยได้นำเอาเทคนิค 10-fold

cross validation และวิธีTrain test มาปรับใช้ในงานวิจัยพัฒนาตัวแบบพยากรณ์ผลผลิตมันสำปะหลังโดยใช้เทคนิคการทำเหมืองข้อมูล และพัฒนาระบบสารสนเทศการพยากรณ์ผลผลิต

มันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์

การออกแบบระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์จะน าเงื่อนไขที่ได้จากต้นไม้ตัดสินใจ ของอัลกอริทึม J48 อัลกอริทึม SimpleCart และอัลกอริทึม LADTree มาพัฒนาเป็นระบบสารสนเทศการพยากรณ์ผลผลิตมันสำปะหลัง การสืบค้นข้อมูล และการประเมินความพึงพอใจผ่านเว็บไซต์ โดยใช้ภาษา PHP ในการพัฒนาเว็บไซต์ ใช้โปรแกรม MySQL ในการจัดการฐานข้อมูล และใช้โปรแกรม Apache เป็นเว็บเซิร์ฟเวอร์ และในส่วนของประเมินความพึงพอใจจากเจ้าหน้าที่สำนักงานเกษตรจังหวัดกำแพงเพชร ผู้ใช้งานทั่วไป และผู้ดูแลระบบ รวม 30 คน มีความพึงพอใจในการใช้งานระบบดังกล่าว

โดยการหาค่าเฉลี่ยจากการตอบแบบประเมินความพึงพอใจผ่านเว็บไซต์คิดเป็น 91% ซึ่งถือว่าอยู่ในระดับที่ดีมาก

ข้อเสนอแนะในการวิจัยพบว่า การสร้างตัวแบบการพยากรณ์ด้วยเทคนิคการจำแนกประเภท (classification) ข้อมูลด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) ไม่สามารถให้คำตอบที่เป็นค่าต่อเนื่องได้ซึ่งจะแสดงเป็นค่าระดับ หรือ ค่าช่วง โดยอาจเลือกใช้วิธีโครงข่ายประสาท (Neural Network) ที่ให้ค่าการพยากรณ์แบบต่อเนื่องได้ และอาจพัฒนาไปเป็นแอปพลิเคชันเพื่อใช้กับสมาร์ทโฟน และแท็บเล็ต เพื่อสะดวกและคล่องตัวในการใช้งานต่อไป