

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในปัจจุบันนี้สภาพเศรษฐกิจของประเทศไทยอยู่ในสภาวะที่มีการแข่งขันสูง ซึ่งมีผลกระทบต่อชีวิตความเป็นอยู่ของประชากร โดยเฉพาะอย่างยิ่งรายได้ซึ่งเป็นปัจจัยสำคัญต่อการดำรงชีวิตของประชากร ทางสำนักงานสถิติแห่งชาติ ได้มีการจัดเก็บข้อมูลการสำรวจด้านการทำงานของประชากร (Labor Forces) โดยเก็บรวบรวมข้อมูลเกี่ยวกับข้อมูลพื้นฐานของประชากร สถานภาพการทำงาน ความพร้อมในการทำงาน การสมัครงาน ความต้องการทำงานเพิ่มเติมจากงานประจำ และรายได้ค่าตอบแทนในรูปแบบต่างๆ โดยได้ทำการจัดเก็บจากประชากรทั่วประเทศ เป็นแบบรายไตรมาส ศูนย์วิจัยมหาวิทยาลัยชิคาโก-มหาวิทยาลัยหอการค้าไทย (UC-UTCC Research Center) ได้จัดเก็บข้อมูลดังกล่าวในฐานข้อมูลปฐมภูมิเพื่ออำนวยความสะดวกให้กับนักวิจัยในการนำข้อมูลมาศึกษาวิจัยต่อไป ดังนั้นทางคณะผู้ดำเนินงานวิจัยจึงมีความประสงค์ในการนำข้อมูลการสำรวจด้านการทำงานของประชากร ประจำปี พ.ศ. 2552 จำนวน 193,176 รายการมาวิเคราะห์โดยใช้เทคนิคเหมืองข้อมูล ในการทำวิจัยเรื่อง “การศึกษาปัจจัยที่มีผลต่อรายได้ของประชากรในประเทศไทย โดยใช้เทคนิคการทำเหมืองข้อมูล”

การทำเหมืองข้อมูล (Data Mining) สามารถแบ่งอัลกอริทึมในการทำงานได้เป็น 2 ประเภทหลัก คือ การสร้างแบบจำลองแบบทำนาย เน้นการจัดกลุ่มโดยอาศัยผลเฉลยที่มีอยู่ซึ่งประกอบด้วย การจำแนกและการถดถอย ส่วนประเภทที่สอง คือการสร้างแบบจำลองเชิงพรรณนาเป็นการหาความสัมพันธ์โดยไม่มีผลเฉลย

งานวิจัยนี้ได้นำเสนอหลักการการทำเหมืองข้อมูลเพื่อสร้างแบบจำลองในการจำแนกประเภทข้อมูล การสำรวจด้านการทำงานของประชากรของสำนักงานสถิติแห่งชาติ ตามประเภทอุตสาหกรรม โดยใช้ต้นไม้การตัดสินใจ (Decision Tree) จากนั้นจะทำการหาปัจจัยที่มีผลต่อรายได้ของประชากรต่อไป

1.2 วัตถุประสงค์ของโครงการวิจัย

1. เพื่อสร้างแบบจำลองในการจำแนกประเภทข้อมูลการสำรวจด้านการทำงานของประชากร ของสำนักงานสถิติแห่งชาติโดยใช้แบบจำลองต้นไม้การตัดสินใจ
2. เพื่อสร้างกฎที่ได้จากแบบจำลองต้นไม้การตัดสินใจ
3. เพื่อหาปัจจัยที่มีผลต่อรายได้ของประชากรในประเทศไทย

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ทราบถึงปัจจัยที่มีผลต่อรายได้ของประชากรของประเทศไทย
2. นำผลที่ได้จากการศึกษาไปใช้ในการวางแผนยุทธศาสตร์เพื่อแก้ปัญหาด้านรายได้ของประชากรในอนาคต

1.4 ขอบเขตของการวิจัย

1. ข้อมูลในการสร้างแบบจำลองต้นไม้การตัดสินใจในงานวิจัยนี้ คือข้อมูลการสำรวจด้านการทำงานของประชากร ของสำนักงานสถิติแห่งชาติ โดยใช้ข้อมูลในปี พ.ศ. 2552 จำนวน 193,176 รายการ โดยจำแนกตามประเภทของอุตสาหกรรม
2. ในการสร้างแบบจำลองต้นไม้การตัดสินใจใช้อัลกอริทึม C4.5 ซึ่งประมวลผลโดยใช้โปรแกรม Weka
3. ทำการทดสอบแบบจำลองต้นไม้การตัดสินใจ โดยนำข้อมูลมาแบ่งเป็น 2 ส่วนคือข้อมูลฝึก (Training Data) และข้อมูลทดสอบ (Testing Data) จนได้ค่าความถูกต้องในระดับที่น่าพอใจ
4. วิเคราะห์ปัจจัยที่มีผลต่อรายได้ของประชากรในประเทศไทย จากกฎที่ได้จากแบบจำลองต้นไม้การตัดสินใจ

1.5 นิยามศัพท์

1. รายได้ คือเงินจากการประกอบอาชีพ แบ่งออกเป็น 2 ประเภท ประเภทแรกคือรายได้ที่เป็นตัวเงิน เช่น ค่าล่วงเวลา โบนัส ประเภทที่สองคือรายได้ที่ไม่ได้เป็นตัวเงิน เช่น รายได้ที่เป็นอาหาร เสื้อผ้าหรือเครื่องแต่งกาย ที่อยู่อาศัย เป็นต้น
2. ประเภทอุตสาหกรรม จากข้อมูลการสำรวจด้านการทำงานของประชากร ของสำนักงานสถิติแห่งชาติ ได้จำแนกประเภทอุตสาหกรรมออกเป็น 17 กลุ่ม คือ

- เกษตรกรรม การล่าสัตว์ และการป่าไม้
- การประมง
- การทำเหมืองแร่และเหมืองหิน
- การผลิต
- การไฟฟ้า ก๊าซ และการประปา
- การก่อสร้าง
- การขายส่ง การขายปลีก การซ่อมแซมยานยนต์ รถจักรยานยนต์ของใช้ส่วนบุคคลและของใช้ในครัวเรือน
- โรงแรมและภัตตาคาร
- การขนส่ง สถานที่เก็บสินค้า และการคมนาคม
- การเป็นตัวกลางทางการเงิน
- กิจกรรมด้านอสังหาริมทรัพย์ การให้เช่าและกิจกรรมทางธุรกิจ
- การบริหารราชการและการป้องกันประเทศ รวมทั้งการประกันสังคมภาคบังคับ
- การศึกษา
- งานด้านสุขภาพและงานสังคมสงเคราะห์
- กิจกรรมด้านการบริการชุมชน สังคมและการบริการส่วนบุคคลอื่นๆ
- ลูกจ้างในครัวเรือนส่วนบุคคล
- องค์การระหว่างประเทศและองค์การต่างประเทศอื่นๆ และสมาชิก

จากการจำแนกประเภทอุตสาหกรรมของสำนักงานสถิติแห่งชาติ พบว่ามีจำนวนประเภทอุตสาหกรรมเป็นจำนวนมาก ซึ่งในรายงานสรุปประจำปี พ.ศ. 2552 ของสำนักงานสถิติแห่งชาติได้สรุปโดยจำแนกประเภทอุตสาหกรรมออกเป็น 3 กลุ่ม คือ ภาคเกษตรกรรม ภาคการผลิต และภาคการบริการและการค้า ทางคณะผู้วิจัยจึงทำการปรับการจำแนกประเภทอุตสาหกรรมเพื่อให้สอดคล้องกับงานวิจัยนี้ โดยแบ่งเป็น 4 ประเภทหลัก ๆ ดังนี้ กลุ่มที่ 1 คือ เกษตรและประมง กลุ่มที่ 2 คือ การผลิต กลุ่มที่ 3 คือ การค้าและการบริการ และกลุ่มที่ 4 คือ อื่น ๆ เช่น ข้าราชการ องค์การระหว่างประเทศ งานด้านสุขภาพและงานสังคมสงเคราะห์ เป็นต้น