การทำนายการปรับตำแหน่งพนักงานโดยใช้การเรียนรู้ของเครื่อง

ณัชชา ชินนาพันธ์[1*], วราภรณ์ วิยานนท์[2]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการทำนายการปรับตำแหน่งพนักงานโดยใช้การเรียนรู้ของเครื่อง เพื่อสร้างแบบจำลองการทำนายการปรับตำแหน่งพนักงาน โดยใช้ชุดข้อมูลการปรับตำแหน่งพนักงานปี 2562 - 2565 จำนวน 212 คน ของฝ่ายบริหารทรัพยากรบุคคล บริษัท ไทยสมุทรประกันชีวิต จำกัด (มหาชน) ผู้วิจัยได้ศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลอง ได้แก่ Support Vector Machine, Decision Tree, Random Forest และ XGBoost ร่วมกับการคัดเลือกฟีเจอร์ (Feature Selection) โดยใช้เทคนิค Recursive Feature Elimination (RFE) และการจัดความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มตัวอย่างเพิ่มกลุ่มน้อยด้วยการสร้างตัวอย่างสังเคราะห์ (Synthetic Minority Over-sampling Technique หรือ SMOTE) จากการทดสอบแบบจำลองที่ได้ประสิทธิภาพสูงสุด คือ แบบจำลอง Support Vector Machines และ แบบจำลอง Random Forest ที่แบ่งข้อมูล 80:20 ร่วมกับการคัดเลือกฟีเจอร์ ได้ค่า Accuracy = 93.02%, Precision = 94.74%, Recall = 97.30% และ F1-Score = 96.00% ที่เท่ากัน จากการทดสอบแบบจำลองที่มีการคัดเลือกฟีเจอร์ก่อนการสร้างแบบจำลองทำให้ได้แบบจำลองที่มีประสิทธิภาพดีกว่าการใช้ฟีเจอร์ทั้งหมดในการสร้างแบบจำลอง การทำ SMOTE สามารถจัดการปัญหาความไม่สมดุลของข้อมูลได้ แต่ไม่ช่วยเพิ่มประสิทธิภาพของแบบจำลอง

คำสำคัญ : การปรับตำแหน่งพนักงาน, การทำนาย, การเรียนรู้ของเครื่อง

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 064-5966966 E-mail address: natcha.chinnapan@g.swu.ac.th

EMPLOYEE PROMOTION PREDICTION USING MACHINE LEARNING

Natcha Chinnapan[1*], Waraporn Viyanon[2]

Abstract

The purpose of this research was to study employee promotion prediction using machine learning. In order to generate a predictive model for employee promotion using dataset of 2019 - 2022 employee position of 212 people from Human Resources Management Division of Ocean Life Insurance Public Company Limited. The researcher has studied and compared the efficiency of models including Support Vector Machine, Decision Tree, Random Forest and XGBoost. Feature selection using Recursive Feature Elimination (RFE) and imbalanced datasets were handled by Synthetic Minority Over-sampling Technique (SMOTE). The best results were Support Vector Machines model and Random Forest model with 80:20 data split with feature selection, Accuracy = 93.02%, Precision = 94.74%, Recall = 97.30% and F1-Score = 96.00%. Model with feature selection before modeling results in a model that performs better than using all features in modeling. SMOTE can address the data asymmetry problem but it does not improve the performance of the model.

Keywords : Employee Promotion, Prediction, Machine Learning

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 064-5966966 E-mail address: natcha.chinnapan@g.swu.ac.th

บทนำ

ทรัพยากรบุคคลถือเป็นปัจจัยหลักปัจจัยหนึ่งที่มีความสำคัญอย่างยิ่งต่อองค์กร เป็นปัจจัยที่ขับเคลื่อนองค์กรให้ดำเนินงานต่างๆ ให้เป็นไปตามเป้าหมายที่ตั้งไว้ให้ประสบความสำเร็จ รวมถึงยังเป็นส่วนสำคัญในการผลักดันกลยุทธ์ขององค์กร และวิสัยทัศน์ของผู้บริหารให้เกิดผลสำเร็จ ปัจจุบันการแข่งขันเพื่อให้ได้มาซึ่งทรัพยากรบุคคลที่มีความสามารถ และการรักษาทรัพยากรบุคคลให้อยู่กับองค์กรอย่างยาวนานที่สุด จึงเป็นเรื่องที่สำคัญ และมีความท้าทายเป็นอย่างมากกับฝ่ายบริหารทรัพยากรบุคคล

การสร้างความก้าวหน้าและการเติบโตในอาชีพการทำงาน เป็นสิ่งที่พนักงานควรได้รับ และเป็นการสร้างความพึงพอใจให้แก่พนักงาน เพื่อดึงดูดให้พนักงานอยู่กับองค์กรต่อไป ซึ่งองค์กรที่มีโครงสร้างการปฏิบัติงานที่ชัดเจน จะมีรายละเอียดของตำแหน่งงานแต่ละตำแหน่งงาน มีแนวทางการปรับตำแหน่งพนักงาน และการประเมินศักยภาพและพฤติกรรมของพนักงานด้านต่างๆ ในแต่ละปี เพื่อเก็บเป็นข้อมูลในการพิจารณาปรับตำแหน่งงาน และประเมินการปรับผลตอบแทนประจำปี ซึ่งฝ่ายบริหารทรัพยากรบุคคลต้องพิจารณาข้อมูลของพนักงานทั้งองค์กรเพื่อคัดกรอง และคัดเลือกพนักงานที่มีความเหมาะสมสำหรับปรับตำแหน่งงาน ดังนั้น งานวิจัยนี้จึงศึกษาการสร้างเทคนิคการเรียนรู้ของเครื่อง เพื่อช่วยในการประกอบการตัดสินใจเลือกพนักงานที่จะปรับตำแหน่ง

จากการศึกษางานวิจัยที่เกี่ยวข้อง มีงานวิจัยเกี่ยวข้องกับการทำนายปรับตำแหน่งพนักงาน มีตัวอย่างดังต่อไปนี้ Kaewwiset et al., 2021 [1] งานวิจัยนี้ทำการศึกษาวิธีการทำนายการปรับตำแหน่งพนักงาน โดยใช้เทคนิคการเรียนรู้ของเครื่อง 3 แบบ คือ Decision tree, Random Forest และ Support Vector Machines โดยทำการแบ่งชุดข้อมูล training set และ test set ที่แตกต่างกัน 3 กลุ่ม คือ 80:20, 70:30 และ 60:40 โดยในแต่ละกลุ่มจะทำ SMOTE กับชุดข้อมูล และไม่ทำ SMOTE ผลลัพธ์จากการศึกษาพบว่าการทำ SMOTE กับชุดข้อมูลให้ความแม่นยำมากกว่าการไม่ทำ SMOTE และการแบ่งชุดข้อมูล training set และ test set ที่ 80:20 ให้ความแม่นยำสูงสุด โดยแบบจำลอง Random Forest เป็นแบบจำลองที่มีประสิทธิภาพสูงสุดที่ 96.32% ซึ่งการทำ SMOTE กับชุดข้อมูล สามารถปรับปรุงความแม่นยำได้สูงสุดถึง 4.80% เมื่อเทียบกับการไม่ทำ SMOTE

Long et al., 2018 [2] งานวิจัยนี้ทำการศึกษาวิธีการทำนายปรับตำแหน่งพนักงาน โดยใช้ข้อมูลของพนักงานองค์กรรัฐวิสาหกิจในประเทศจีน ใช้เทคนิคการเรียนรู้ของเครื่อง 6 แบบ คือ K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF) และ Adaboost โดยจะมีการดำเนินการ 3 หลัก คือ ขั้นตอนที่ 1 สร้างฟีเจอร์ที่เป็นข้อมูลพื้นฐานส่วนบุคคล และข้อมูลตามกลยุทธ์ 5 ประการ ได้แก่ กลยุทธ์ 1 คือ เลือกค่าที่ไม่ซ้ำ, กลยุทธ์ 2 คือ เลือกโหมด, กลยุทธ์ 3 คือ เลือกค่าสูงสุดหรือต่ำสุด, กลยุทธ์ 4 คือ นับจำนวนค่าต่างๆ และกลยุทธ์ 5 คือ คำนวณความแตกต่างระหว่างสองวัน ขั้นตอนที่ 2 ตรวจสอบฟีเจอร์ และวิเคราะห์ความสัมพันธ์ (correlation) ระหว่างฟีเจอร์ ขั้นตอนที่ 3 สร้างแบบจำลอง ทดสอบแบบจำลอง และวิเคราะห์ความสำคัญของฟีเจอร์ ผลลัพธ์จากการศึกษาพบว่าแบบจำลอง Random Forest เป็นแบบจำลองที่มีประสิทธิภาพสูงสุดจาก 6 แบบจำลอง ได้ค่า accuracy อยู่ที่ 90% เมื่อการตรวจสอบความถูกต้องของฟีเจอร์ต่างๆ คำนวณค่า Gini ของแต่ละฟีเจอร์เพื่อวิเคราะห์เพิ่มเติม พบว่าฟีเจอร์ข้อมูลตามกลยุทธ์ 5 ประการมีผลกระทบกับการปรับตำแหน่งมากกว่าเมื่อเทียบกับฟีเจอร์ข้อมูลพื้นฐานส่วนบุคคล และจากฟีเจอร์ทั้งหมด ซึ่งจำนวนปีการทำงาน ตำแหน่งที่แตกต่างกัน และระดับแผนกสูงสุดส่งผลกระทบอย่างมากกับการปรับตำแหน่ง

Alqahtani & Almaleh, 2022 [3] งานวิจัยนี้ทำการศึกษาวิธีการทำนายการปรับตำแหน่งพนักงาน และหาฟีเจอร์ที่สำคัญที่ส่งผลต่อการปรับตำแหน่ง โดยใช้ชุดข้อมูลเปิดจาก Kaggle ปี 2020 ที่ประกอบข้อมูลบริษัทข้ามชาติ 54,808 แถว และ 13

คอลัมน์ ซึ่งเป็นชุดข้อมูลที่ครอบคลุมธุรกิจ 9 ประเภท ใช้เทคนิคการเรียนรู้ของเครื่อง 6 แบบ คือ K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine และ Ensemble models (Adaboosting and Gradient Boosting models) ผลลัพธ์จากการศึกษาพบว่าแบบจำลอง Gradient Boosting มีประสิทธิภาพดีกว่าแบบจำลองประเภทอื่น ได้ค่า accuracy อยู่ที่ 93.90% และฟีเจอร์ที่สำคัญที่สุดในการทำนายการปรับตำแหน่ง คือ Previous Year Ratings Feature

วิธีดำเนินการ

ขั้นตอนที่ 1 : การกำหนดประชากรและกลุ่มตัวอย่าง

งานวิจัยนี้ใช้ข้อมูลการปรับตำแหน่งพนักงานปี 2562 - 2565 จำนวน 212 คน ของฝ่ายบริหารทรัพยากรบุคคล บริษัท ไทยสมุทรประกันชีวิต จำกัด (มหาชน) ที่มีการระบุว่าผ่านการปรับตำแหน่ง หรือไม่ผ่านการปรับตำแหน่ง ซึ่งภายในชุดข้อมูลประกอบด้วยข้อมูลตัวแปรทั้งหมด 14 คอลัมน์ และมีจำนวนแถวทั้งหมด 212 แถว

ขั้นตอนที่ 2 : การเตรียมข้อมูลและการวิเคราะห์ข้อมูล

เริ่มต้นจากการนำเข้าชุดข้อมูลการปรับตำแหน่งพนักงานเพื่อมาทำการแสดงรายละเอียด จากนั้นทำการเตรียมข้อมูล หรือ Data Preprocessing เพื่อทำให้ข้อมูลอยู่ในรูปแบบที่ง่ายต่อการนำไปใช้งาน และวิเคราะห์ข้อมูล โดยจะทำการเปลี่ยนชื่อคอลัมน์, เปลี่ยนชนิดข้อมูล, ตรวจสอบค่าว่าง และข้อมูลที่ซ้ำกันในแต่ละแถว จากนั้นทำการวิเคราะห์ข้อมูลต่างๆ เพื่อค้นหาข้อมูลที่สำคัญที่แฝงอยู่ในกลุ่มข้อมูลนั้น โดยจะทำการหารูปแบบความเชื่อมโยงระหว่างกัน เริ่มจากการหาค่าทางสถิติเบื้องต้น และทำการเรียกดูข้อมูลจากคอลัมน์ต่างๆ ว่ามีตัวแปรแต่ละตัวแปรเป็นจำนวนเท่าไหร่ และสัดส่วนของแต่ละตัวแปรกับผลการปรับตำแหน่งว่าเป็นจำนวนเท่าใด เพื่อหาว่าข้อมูลมีความเป็นไปอย่างไร จากนั้นทำการแปลงข้อมูลโดยใช้เทคนิค Label encoder เพื่อแปลงข้อมูลให้เป็นตัวเลข เมื่อแปลงข้อมูลให้เป็นตัวเลขแล้ว จึงหาค่าความสัมพันธ์ของข้อมูล (Correlation) เพื่อตรวจสอบว่าฟีเจอร์ใดมีความสัมพันธ์กับจัดกลุ่มการปรับตำแหน่งมากที่สุด

ขั้นตอนที่ 3 : การแบ่งชุดข้อมูลและการปรับช่วงขอบเขตฟีเจอร์

เมื่อทำการเตรียมข้อมูลเรียบร้อยแล้ว จึงนำชุดข้อมูลไปแบ่งเป็น 2 ส่วน คือ ชุดข้อมูลสำหรับสร้างแบบจำลอง (Training Set) และชุดข้อมูลสำหรับทดสอบแบบจำลอง (Test Set) ด้วยเทคนิค Train-test split ในอัตราส่วน 2 อัตราส่วน คือ 70:30 และ 80:20 โดยอัตราส่วน 70:30 จะมีข้อมูลสำหรับสร้างแบบจำลองทั้งหมด 148 ข้อมูล และข้อมูลสำหรับทดสอบแบบจำลองทั้งหมด 64 ข้อมูล อัตราส่วน 80 : 20 จะมีข้อมูลสำหรับสร้างแบบจำลองทั้งหมด 169 ข้อมูล และข้อมูลสำหรับทดสอบแบบจำลองทั้งหมด 23 ข้อมูล โดยการเตรียมข้อมูลต่อไปจะใช้กับชุดข้อมูลสำหรับสร้างแบบจำลองเท่านั้น จากนั้นทำการปรับช่วงขอบเขตของฟีเจอร์ เนื่องจากช่วงข้อมูลของฟีเจอร์มีความหลากหลาย ซึ่งจะใช้เทคนิค StandardScaler [4] ที่เป็นการปรับค่าข้อมูลตัวเลขโดยคำนวณจากค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐาน เพื่อให้ข้อมูลตัวเลขอยู่ภายในขอบเขตเดียวกัน

ขั้นตอนที่ 4 : การคัดเลือกฟีเจอร์และการจัดการความไม่สมดุลของข้อมูล

ในงานวิจัยนี้ใช้การคัดเลือกฟีเจอร์แบบ Wrapper Methods โดยใช้เทคนิค Recursive Feature Elimination (RFE) [5] กับแบบจำลองทั้ง 4 แบบจำลอง ได้แก่ Support Vector Machine [6], Decision Tree [7], Random Forest และ XGBoost

[8] เพื่อหาฟีเจอร์ที่มีความสำคัญกับประสิทธิภาพของแต่ละแบบจำลอง โดยจะมีการจัดให้อันดับให้กับฟีเจอร์แต่ละตัว และทำการหาค่า Feature Importance ของแต่ละฟีเจอร์ จากนั้นจัดการความสมดุลของข้อมูล โดยทำให้ข้อมูลพนักงานผ่านและไม่ผ่านการปรับตำแหน่งมีจำนวนเท่ากัน โดยใช้เทคนิค Synthetic Minority Over-sampling Technique หรือ SMOTE [9] โดยกำหนดค่า K-neighbors = 5 หลังการทำ SMOTE ทำให้ข้อมูลผลการปรับตำแหน่งมีจำนวนคนผ่านการปรับตำแหน่ง คือ 1 และไม่ผ่านการปรับตำแหน่ง คือ 0 เท่ากัน โดยชุดข้อมูลในอัตราส่วน 70:30 มีข้อมูลทั้ง 2 กลุ่มเท่ากันที่ 128 คน รวม 256 ข้อมูล และอัตราส่วน 80:20 มีข้อมูลทั้ง 2 กลุ่มเท่ากันที่ 146 คน รวม 292 ข้อมูล ที่จะใช้เป็นชุดข้อมูลสำหรับสร้างแบบจำลอง

ขั้นตอนที่ 4 : การสร้างแบบจำลอง และการประเมินประสิทธิภาพแบบจำลอง

แบบจำลองที่สร้างในงานวิจัยนี้ มี 4 แบบจำลอง ได้แก่ Support Vector Machine, Decision Tree, Random Forest และ XGBoost ในการสร้างแบบจำลองจะทำการปรับจูนพารามิเตอร์ (Hyperparameter Tuning) [10] ร่วมด้วย โดยจะใช้เทคนิค GridSearchCV เพื่อค้นหาค่า Hyperparameter จากค่าที่กำหนดไว้แล้ว และเลือกค่าที่ทำให้แบบจำลองมีประสิทธิภาพสูงที่สุด เมื่อแบบจำลองทำการเรียนรู้จากชุดข้อมูลสำหรับสร้างแบบจำลองเรียบร้อยแล้ว จึงนำชุดข้อมูลสำหรับทดสอบแบบจำลองมาใช้ในการวัดประสิทธิภาพของแบบจำลองโดยใช้ตาราง Confusion Matrix ได้แก่ Accuracy, Precision, Recall และ F1 Score และทำการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ฟีเจอร์ทั้งหมด แบบจำลองที่มีการคัดเลือกฟีเจอร์ แบบจำลองที่มีการจัดการความไม่สมดุลของข้อมูล และแบบจำลองที่มีการคัดเลือกฟีเจอร์และการจัดการความไม่สมดุลของข้อมูลร่วมด้วย

ผลการวิจัยและอภิปรายผลการวิจัย

1. การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง

จากตารางที่ 1-4 แสดงให้เห็นว่าแบบจำลองที่มีประสิทธิภาพสูงสุด คือ แบบจำลอง Support Vector Machines และ Random Forest ที่แบ่งข้อมูล 80:20 ร่วมกับการคัดเลือกฟีเจอร์ให้ประสิทธิภาพสูงสุดเท่ากัน คือ Accuracy 93.02%, Precision 94.74%, Recall 97.30% และ F1-Score 96.00%

ตารางที่ 1 แสดงประสิทธิภาพการทดสอบจากแบบจำลอง Support Vector Machine

แบบจำลอง	อัตราส่วนข้อมูล	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine without SMOTE and Feature Selection	70:30	89.06	91.38	96.36	93.81
	80:20	86.05	89.74	94.59	92.11
Support Vector Machine with SMOTE	70:30	85.94	85.94	100	92.44
	80:20	88.37	88.1	100	93.67
Support Vector Machine with Feature Selection	70:30	89.06	91.38	96.36	93.81
	80:20	93.02	94.74	97.30	96.00
Support Vector Machine with SMOTE and Feature Selection	70:30	75.00	95.35	74.55	83.67
	80:20	72.09	90.32	75.68	82.35

ตารางที่ 2 แสดงประสิทธิภาพการทดสอบจากแบบจำลอง Decision Tree

แบบจำลอง	อัตราส่วนข้อมูล	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree without SMOTE and Feature Selection	70:30	84.38	92.45	89.09	90.74
	80:20	90.70	90.24	100	94.87
Decision Tree with SMOTE	70:30	81.25	92.16	85.45	88.68
	80:20	79.09	93.75	81.08	86.96
Decision Tree with Feature Selection	70:30	85.94	94.23	89.09	91.59
	80:20	88.37	90	97.3	93.51
Decision Tree with SMOTE and Feature Selection	70:30	70.31	97.37	67.27	79.57
	80:20	72.09	96.3	70.27	81.25

ตารางที่ 3 แสดงประสิทธิภาพการทดสอบจากแบบจำลอง Random Forest

แบบจำลอง	อัตราส่วนข้อมูล	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest without SMOTE and Feature Selection	70:30	90.62	92.98	96.36	94.64
	80:20	90.70	92.31	97.3	94.74
Random Forest with SMOTE	70:30	85.94	96.00	87.27	91.43
	80:20	83.72	94.12	86.49	90.14
Random Forest with Feature Selection	70:30	89.06	92.86	94.55	93.69
	80:20	93.02	94.74	97.30	96.00
Random Forest with SMOTE and Feature Selection	70:30	67.19	97.22	63.64	76.92
	80:20	74.42	96.43	72.97	83.08

ตารางที่ 4 แสดงประสิทธิภาพการทดสอบจากแบบจำลอง Random Forest

แบบจำลอง	อัตราส่วนข้อมูล	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost without SMOTE and Feature Selection	70:30	89.06	94.44	92.73	93.58
	80:20	86.05	91.89	91.89	91.89
XGBoost with SMOTE	70:30	78.12	95.56	78.18	86.00
	80:20	76.74	96.55	75.68	84.85
XGBoost with Feature Selection	70:30	89.06	92.86	94.55	93.69
	80:20	88.37	94.44	91.89	93.15
XGBoost with SMOTE and Feature Selection	70:30	62.50	94.29	60.00	73.00
	80:20	72.09	96.30	70.27	81.25

2. การเปรียบเทียบภาพของแบบจำลองที่ทำการคัดเลือกฟีเจอร์กับประสิทธิภาพของแบบจำลองที่ใช้ฟีเจอร์ทั้งหมด

จากตารางที่ 5 เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ทำการคัดเลือกฟีเจอร์ก่อนการสร้างแบบจำลองกับประสิทธิภาพของแบบจำลองที่ใช้ฟีเจอร์ทั้งหมด ผลลัพธ์ คือ แบบจำลองที่ทำการคัด เลือกฟีเจอร์ที่มีประสิทธิภาพสูงสุด คือ แบบจำลอง Support Vector Machines และ Random Forest ที่แบ่งข้อมูล 80:20 มีประสิทธิภาพสูงสุดเท่ากัน คือ Accuracy 93.02%, Precision 94.74%, Recall 97.30% และ F1-Score 96.00% ซึ่งเป็นแบบจำลองที่มีประสิทธิภาพสูงสุดในการทดสอบ และแบบจำลองที่ใช้ฟีเจอร์ทั้งหมดที่มีประสิทธิภาพสูงสุด คือ Decision Tree ที่แบ่งข้อมูล 80:20 ได้ค่า Accuracy 90.70%, Precision 90.24 %, Recall 100% และ F1-Score 94.87% ดังนั้น การคัดเลือกฟีเจอร์ก่อนการสร้างแบบจำลองให้ประสิทธิภาพที่ดีกว่าการใช้ฟีเจอร์ทั้งหมดในการสร้างแบบจำลอง ยกเว้นแบบจำลอง Decision Tree ที่แบ่งข้อมูล 80:20 เพียงแบบจำลองเดียวที่ใช้ฟีเจอร์ทั้งหมดแต่ได้ประสิทธิภาพดีกว่าแบบจำลอง Decision Tree ที่แบ่งข้อมูล 80:20 ร่วมกับการคัดเลือกฟีเจอร์

ตารางที่ 5 แสดงการเปรียบเทียบประสิทธิภาพการทดสอบแบบจำลองเมื่อคัดเลือกฟีเจอร์กับแบบจำลองที่ใช้ฟีเจอร์ทั้งหมด

แบบจำลอง	อัตราส่วนข้อมูล	All Feature				Feature Selection			
		Accuracy	Percision	Recall	F1-Score	Accuracy	Percision	Recall	F1-Score
Support Vector Machines	70:30	89.06	91.38	96.36	93.81	89.06	91.38	96.36	93.81
	80:20	86.05	89.74	94.59	92.11	93.02	94.74	97.30	96.00
Decision Tree	70:30	84.38	92.45	89.09	90.74	85.94	94.23	89.09	91.59
	80:20	90.70	90.24	100	94.87	88.37	90.00	97.30	93.51
Random Forest	70:30	90.62	92.98	96.36	94.64	89.06	92.86	94.55	93.69
	80:20	90.70	92.31	97.30	94.74	93.02	94.74	97.30	96.00
XGBoost	70:30	89.06	94.44	92.73	93.58	89.06	92.86	94.55	93.69
	80:20	86.05	91.89	91.89	91.89	88.37	94.44	91.89	93.15

3. การเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลที่ทำ SMOTE และไม่ทำ SMOTE

ตารางที่ 6 แสดงการเปรียบเทียบประสิทธิภาพการทดสอบแบบจำลองใช้ข้อมูลที่ทำ SMOTE และไม่ทำ SMOTE

แบบจำลอง	อัตราส่วนข้อมูล	without SMOTE				with SMOTE			
		Accuracy	Percision	Recall	F1-Score	Accuracy	Percision	Recall	F1-Score
Support Vector Machines	70:30	89.06	91.38	96.36	93.81	85.94	85.94	100	92.44
	80:20	86.05	89.74	94.59	92.11	88.37	88.10	100	93.67
Decision Tree	70:30	84.38	92.45	89.09	90.74	81.25	92.16	85.45	88.68
	80:20	90.70	90.24	100	94.87	79.07	93.75	81.08	86.96
Random Forest	70:30	90.62	92.98	96.36	94.64	85.94	96.00	87.27	91.43
	80:20	90.70	92.31	97.30	94.74	83.72	94.12	86.49	90.14
XGBoost	70:30	89.06	94.44	92.73	93.58	78.12	95.56	78.18	86.00
	80:20	86.05	91.89	91.89	91.89	76.74	96.55	75.68	84.85

จากตารางที่ 6 สำหรับแบบจำลองที่ทำการจัดการความสมดุลของข้อมูลด้วยเทคนิค SMOTE เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลที่ทำ SMOTE และไม่ทำ SMOTE ผลลัพธ์ คือ แบบจำลองที่ใช้ข้อมูลที่ทำ SMOTE ที่มีประสิทธิภาพสูงสุด คือ แบบจำลอง Support Vector Machines ที่แบ่งข้อมูล 80:20 ได้ค่า Accuracy 88.37%, Precision 88.10%, Recall 100% และ F1-Score 93.67% ซึ่งมีประสิทธิภาพมากกว่าเมื่อเทียบแบบจำลอง Support Vector Machines ที่แบ่งข้อมูล 80:20 ที่ไม่ทำ SMOTE แต่เมื่อเปรียบเทียบประสิทธิภาพของทุกแบบจำลองผลลัพธ์ คือ แบบจำลองที่ใช้การจัดการความสมดุลของข้อมูลด้วยเทคนิค SMOTE ได้ประสิทธิภาพที่ต่ำกว่าการไม่ทำ SMOTE ร่วมด้วยในทุกแบบจำลอง

สรุปผลการวิจัย

งานวิจัยนี้ศึกษาการทำนายผลการปรับตำแหน่งงาน โดยใช้ข้อมูลการปรับตำแหน่งพนักงาน ของฝ่ายบริหารทรัพยากรบุคคล บริษัท ไทยสมุทรประกันชีวิต จำกัด (มหาชน) มาใช้ในการสร้างแบบจำลองทั้งหมด 4 แบบจำลอง ได้แก่ Support Vector Machine, Decision Tree, Random Forest และ XGBoost ร่วมกับการคัดเลือกฟีเจอร์ด้วยเทคนิค RFE และการจัดการความไม่สมดุลของข้อมูลด้วยเทคนิค SMOTE โดยในการสร้างแบบจำลองได้ทำการปรับค่าพารามิเตอร์ด้วยเทคนิค GridSearchCV ซึ่งแบบจำลองที่ได้ประสิทธิภาพสูงสุด คือ แบบจำลอง Support Vector Machines และ แบบจำลอง Random Forest ที่แบ่งข้อมูล 80:20 ร่วมกับการคัดเลือกฟีเจอร์ ได้ค่า Accuracy 93.02%, Precision 94.74%, Recall 97.30% และ F1-Score 96.00% ซึ่งการคัดเลือกฟีเจอร์ก่อนการสร้างแบบจำลองทำให้ได้แบบจำลองที่มีประสิทธิภาพดีกว่าการใช้ฟีเจอร์ทั้งหมดในการสร้างแบบจำลอง การทำ SMOTE สามารถจัดการปัญหาความไม่สมดุลของข้อมูลได้ แต่ไม่ช่วยเพิ่มประสิทธิภาพของแบบจำลอง และทำให้แบบจำลองมีประสิทธิภาพต่ำกว่าเดิม โดยเฉพาะกับแบบจำลองที่มีความซับซ้อนสูงอย่าง XGBoost และการพิจารณาฟีเจอร์พบว่าฟีเจอร์ที่อยู่ในกลุ่มข้อมูลคะแนนการประเมินผลงานและพฤติกรรม เป็นฟีเจอร์ที่มีความสำคัญมากกว่าฟีเจอร์ที่อยู่กลุ่มข้อมูลพนักงาน

งานวิจัยในอนาคต ผู้วิจัยเล็งเห็นจากปัญหาที่พบในการดำเนินการวิจัยครั้งนี้ทั้งหมด เห็นควรว่าในขั้นตอนการเตรียมข้อมูล เปลี่ยนการแปลงข้อมูลเป็นตัวเลขโดยใช้เทคนิค One Hot Encoder ซึ่งเป็นเทคนิคที่เหมาะกับข้อมูลหมวดหมู่หรือประเภทที่ไม่สามารถเรียงลำดับได้อย่างชัดเจน ซึ่งเหมาะกับข้อมูลที่ใช้ในงานวิจัยนี้ และจัดการความไม่สมดุลของข้อมูลโดยใช้เทคนิค Random Over-sampling เพิ่มเติม เพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพมากขึ้นกับชุดข้อมูล รวมถึงศึกษาแบบจำลองอื่นที่อาจจะสามารถเรียนรู้และทำนายออกมาได้มีความแม่นยำและมีประสิทธิภาพที่ดีกว่า เช่น AdaBoost, Gradient Boosting

กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยการให้ความช่วยเหลือแนะนำของ ผศ.ดร.วราภรณ์ วิยานนท์ ซึ่งเป็นอาจารย์ที่ปรึกษา ที่ให้คำแนะนำ ข้อคิดเห็น ตรวจสอบ และแก้ไขร่างสารนิพนธ์มาโดยตลอด และขอขอบคุณคณาจารย์ทุกท่านในสาขาวิชาวิทยาการข้อมูล มหาวิทยาลัยศรีนครินทรวิโรฒ และความอนุเคราะห์ข้อมูลที่ใช้ในการดำเนินการวิจัย จากฝ่ายบริหารทรัพยากรบุคคล บริษัท ไทยสมุทรประกันชีวิต จำกัด (มหาชน) รวมถึงการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] Kaewwiset, T., Temdee, P., & Yooyativong, T. (2021). Employee classification for personalized professional training using machine learning techniques and SMOTE. 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering.

[2] Long, Y., Liu, J., Fang, M., Wang, T., & Jiang, W. (2018). Prediction of employee promotion based on personal basic features and post features. Proceedings of the International Conference on Data Processing and Applications.

[3] Alqahtani, F. A., & Almaleh, A. (2022). Analysis and Prediction of Employee Promotions Using Machine Learning. 2022 5th International Conference on Data Science and Information Technology (DSIT)(Zheng & Casari, 2018)

[4] Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.".

[5] Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.

[6] Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).

[7] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.

[8] TITIPATA. (2018). [ML] Bagging หรือ Boosting คืออะไร ทำงานอย่างไร? https://tupleblog.github.io/bagging-boosting/

[9] Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR), 5(4), 1-29.

[10] Owen, L. (2022). Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning. Packt Publishing. https://books.google.co.th/books?id=CqF-EAAAQBAJ