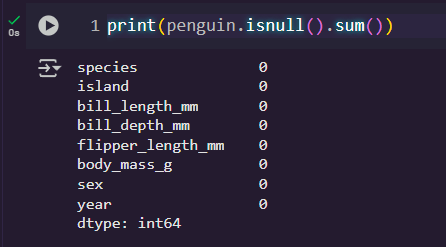
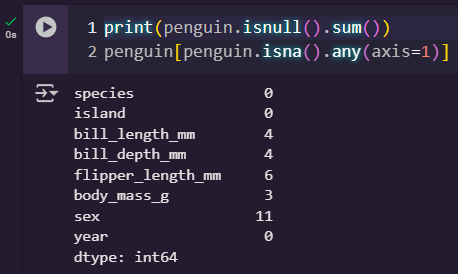
**รายงานผลการทดสอบโมเดลการจำแนกประเภท  
(Classification Model Performance Report)**

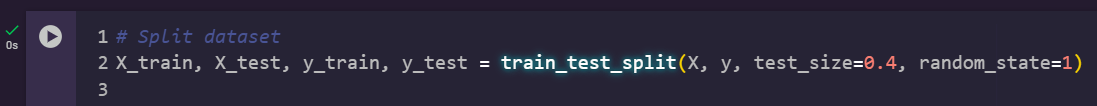
**ชื่อผู้จัดทำ:**   ภาณุทัศ ศิริกุล  **รหัสนักศึกษา:**   650510714     
**กระบวนวิชา:**   204423 - Data Mining    
**ภาคการศึกษา:**   2  **ชื่อไฟล์ Dataset:**   penguin.csv  

**1. รายละเอียดข้อมูล (Dataset Description)  
 ข้อมูลในไฟล์ Dataset มีจำนวนคุณลักษณะ (Attribute):** 6 คุณลักษณะ ได้แก่ ['island', 'bill\_length\_mm', 'bill\_depth\_mm', 'flipper\_length\_mm', 'body\_mass\_g', 'sex'] โดยคุณลักษณะ ปี 'year' ไม่ได้ถูกนำมาใช้เป็นคุณลักษณะในการจำแนกประเภท **รายชื่อ Class ใน Dataset:** 3 ชั้นข้อมูล (Multi-class Classification) ได้แก่ ('Adelie', 'Gentoo', 'Chinstrap') อยู่ในคุณลักษณะ ['species'] **จำนวนข้อมูลในแต่ละ Class:**   
 Class 1 ('Adelie') มี 152 ตัวอย่าง Class 2 ('Gentoo') มี 124 ตัวอย่าง  
 Class 3 ('Chinstrap') มี 68 ตัวอย่าง

**2. การเตรียมข้อมูล (Data Preprocessing)  
 ความจำเป็นในการทำ Data Preprocessing:** ชุดข้อมูล penguin ควรทำการ Data Preprocessing เนื่องจากชุดข้อมูลมีการตรวจพบข้อมูลที่สูญหายใน 17 ระเบียน โดย 'bill\_length\_mm' สูญหาย 4 ตัวอย่าง, 'bill\_depth\_mm' สูญหาย 4 ตัวอย่าง, 'flipper\_length\_mm' สูญหาย 4 ตัวอย่าง, 'body\_mass\_g' สูญหาย 3 ตัวอย่าง, 'sex' สูญหาย 11 ตัวอย่าง เนื่องจากชุดข้อมูลมีการกระจายตัวที่ค่อนข้างดี ไม่ Imbalance มากเกินไป ดังนั้นสิ่งที่ต้องทำจึงมีแค่การทำ Data Cleaning กำจัดข้อมูลสูญหายด้วยวิธีการแทนค่า Imputation  
 **ขั้นตอนการทำ Data Cleaning:** ขั้นตอนที่ 1:ทำการหาค่า Missing Values ว่ามีหรือไม่ ถ้ามีให้ระบุคุณลักษณะที่พบค่าสูญหาย ขั้นตอนที่ 2:จัดการค่าสูญหาย สำหรับคุณลักษณะข้อมูลเชิงปริมาณใช้วิธี Mean Imputation  
 ขั้นตอนที่ 3:จัดการค่าสูญหายสำหรับคุณลักษณะข้อมูลเชิงคุณภาพใช้วิธี One-Hot Encoding โดยเฉพาะคุณลักษณะ ['island', 'sex'] ซึ่งไม่ใช่ข้อมูลที่เป็นตัวเลขได้ออกมาเป็น 5 คุณลักษณะ ดังนี้ ['island\_Biscoe', 'island\_Dream', 'island\_Torgersen', 'sex\_female', 'sex\_male']  
 **ผลลัพธ์จากการทำ Data Preprocessing (Code):** ผลลัพธ์ก่อนและหลังจากการทำ Data Cleaning

****

**3. การแบ่งชุดข้อมูล (Data Splitting)** – Code การแบ่งข้อมูลเป็น Training และ Test Set (60:40)

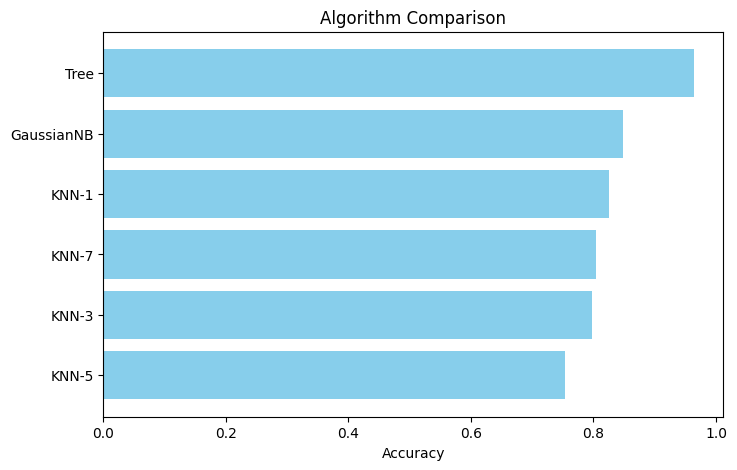


**4. การทดสอบโมเดล (Model Evaluation)  
 โมเดลที่เลือกจากการเรียน:** K-Nearest Neighbor (KNN), Decision Tree, Gaussian Naive Bayes  
 **โมเดลที่เพิ่มเข้ามา:** Gradient Boosting Classifier (GBC) **รายงานผลลัพธ์จากโมเดล Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decision Tree** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Adelie** | **0.98** | **0.95** | **0.97** | **60** |
| **Chinstrap** | **0.92** | **0.96** | **0.94** | **25** |
| **Gentoo** | **0.98** | **1.00** | **0.99** | **53** |
| **Accuracy** | **-** | **-** | **0.97** | **138** |
| **Macro Avg** | **0.96** | **0.97** | **0.97** | **138** |
| **Weighted Avg** | **0.97** | **0.97** | **0.97** | **138** |

จากการทดลองพบว่าประสิทธิภาพของโมเดลการจำแนกประเภท Decision Tree ด้วยวิธี Split Dataset (60:40) นั้นให้ผลลัพธ์การทำนายในช่วง Training Model ที่มีความแม่นยำสูงมาก

**กราฟแสดงค่าความแม่นยำในแต่ละโมเดลที่เลือกจากการเรียน**

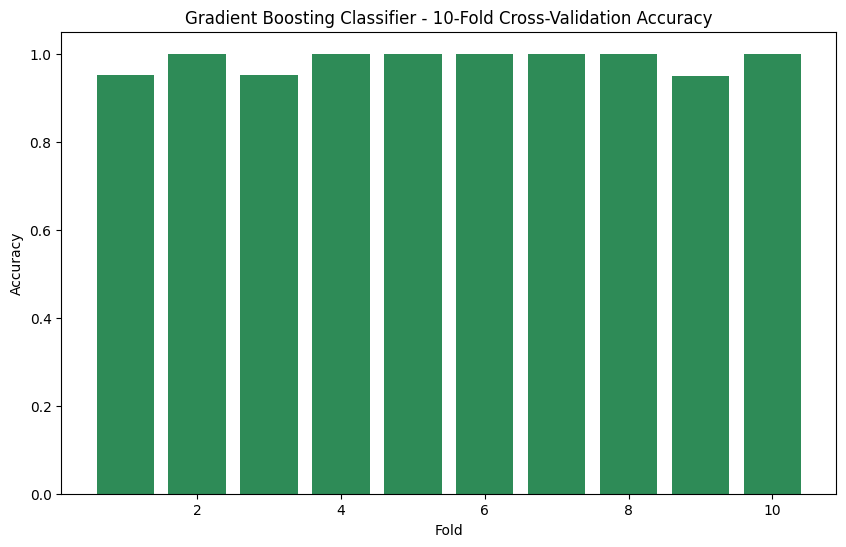


**รายงานผลลัพธ์จากโมเดล GBC (10-Fold Cross-Validation)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GBC (Test Set)** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Adelie** | **0.98** | **1.00** | **0.99** | **60** |
| **Chinstrap** | **1.00** | **0.96** | **0.98** | **25** |
| **Gentoo** | **1.00** | **1.00** | **1.00** | **53** |
| **Accuracy** | **-** | **-** | **0.99** | **138** |
| **Macro Avg** | **0.99** | **0.99** | **0.99** | **138** |
| **Weighted Avg** | **0.99** | **0.99** | **0.99** | **138** |

จากการทดลองพบว่าประสิทธิภาพของโมเดลการจำแนกประเภท Gradient Boosting ด้วยวิธี 10-Fold Cross-Validation นั้นให้ผลลัพธ์การทำนายในช่วง Training Model ที่มีความแม่นยำสูงมากกว่า Decision Tree

**กราฟแสดงค่าความแม่นยำในแต่ละ Fold ของโมเดล Gradient Boosting (10-Fold)**

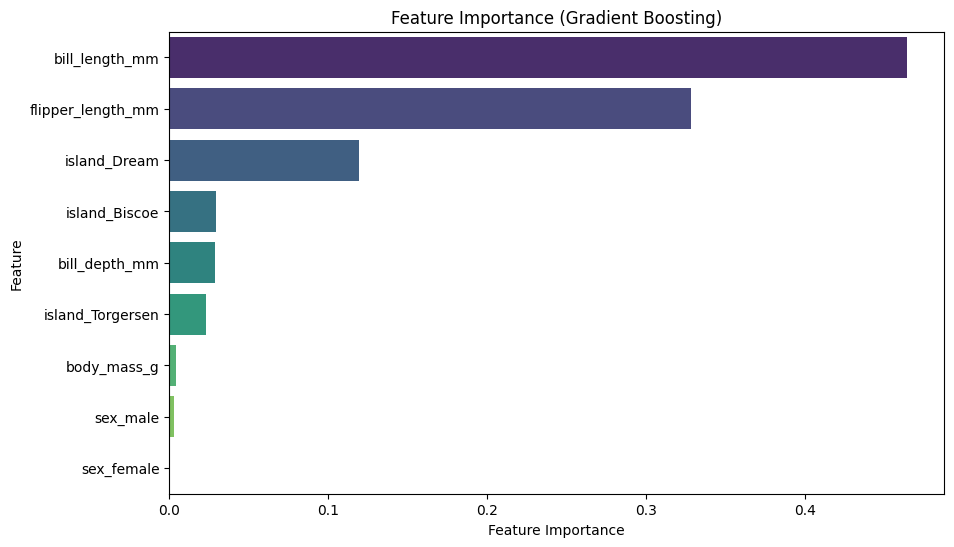


**5. รายงานผลประสิทธิภาพของการทดสอบโมเดลที่เลือก (Gradient Boosting) กับชุดข้อมูล Test Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GBC (Test Set)** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Adelie** | **0.98** | **1.00** | **0.99** | **60** |
| **Chinstrap** | **1.00** | **0.96** | **0.98** | **25** |
| **Gentoo** | **1.00** | **1.00** | **1.00** | **53** |
| **Accuracy** | **-** | **-** | **0.99** | **138** |
| **Macro Avg** | **0.99** | **0.99** | **0.99** | **138** |
| **Weighted Avg** | **0.99** | **0.99** | **0.99** | **138** |

จากการทดลองพบว่าประสิทธิภาพของโมเดลการจำแนกประเภท Gradient Boosting ซึ่งเป็นโมเดลที่ใช้ ensemble learning แบบ boosting นั้น สามารถจำแนกประเภทของ Specie นกเพนกวินได้แม่นยำมากกว่าโมเดล Decision Tree หลังจากที่นำไปทดสอบกับชุดข้อมูล Test Dataset ถึงแม้ว่าโมเดล Decision Tree จะมีความแม่นยำมากกว่า K-Nearest Neighbor (KNN) และ Gaussian Naive Bayes ก็ตาม

**กราฟแสดงความสำคัญในแต่ละคุณลักษณะของโมเดล Gradient Boosting**



**6. สรุปผลการทดลอง (Experiment Conclusion)** จากการทดลองการจำแนกประเภท Species ของนกเพนกวิน พบว่าโมเดลที่มีความแม่นยำมากที่สุด คือ Gradient Boosting และโมเดลที่มีความแม่นยำสูงสุดรองลงมา ได้แก่ Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbor (KNN) ที่ K=1, K=7, K=3, และ K=5 ตามลำดับ

**7. ถ้าต้องเลือก Model เดียว เพื่อนำ Model ไปใช้ นักศึกษาจะเลือก Model ใด เพราะอะไร** เลือกโมเดล Gradient Boosting Classifier เนื่องจากเป็นโมเดลประเภท Ensemble ที่ทำการรวม Weak Learner หลายตัว ทำให้โมเดลมีความแม่นยำโดยทั่วไปสูงกว่าโมเดลที่ไม่ใช่ Ensemble ดังนั้น Gradient Boosting จึงมีความเหมาะสมที่จะนำไปใช้ต่อมากกว่าโมเดลอื่น โดยเฉพาะการนำไปใช้จำแนกประเภทกับ Unseen Data ซึ่งโมเดล GBC จะสามารถทำนายได้โดยที่ความเสี่ยงต่อการเกิด Overfitting นั้นมีน้อยกว่า