K-Nearest Neighbors

6440126922 นิปุณ อังควิชัย

จากการสร้างแบบโมเดล KNN ทำนาย Breast Cancer ตามขั้นตอนใน Google Colab notebook knn.ipynb โดยจากข้อมูลมี Label distribution ใน dataset ที่แบ่งไว้เทรนและทดสอบดังนี้ เมื่อแบ่ง Dataset เป็น Training set, Dev set และ Test set ในสัดส่วน 70:10:20

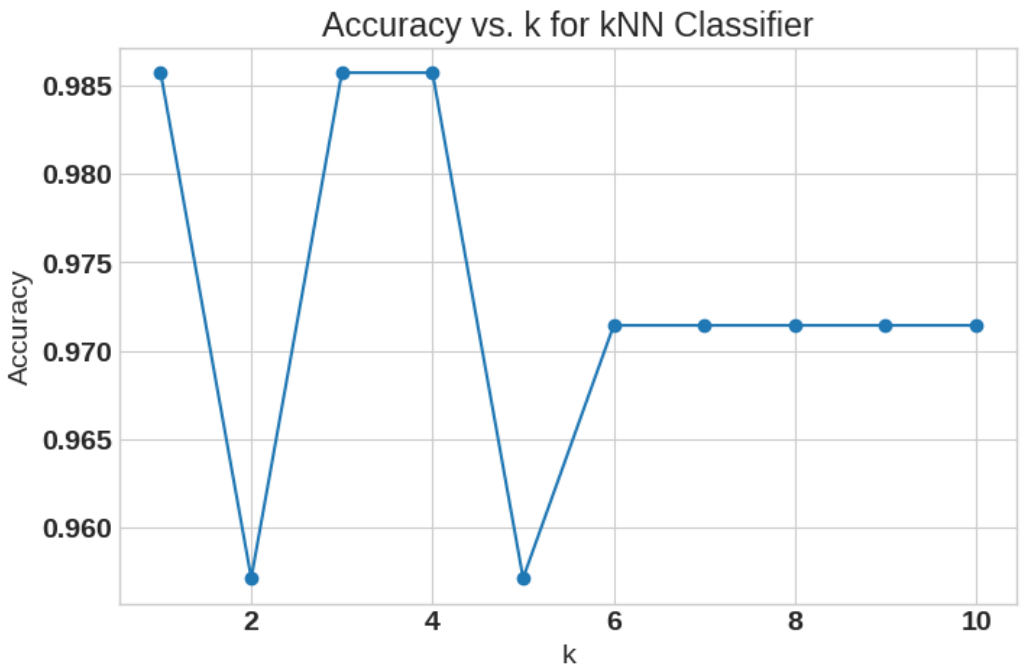
	Positive label (1)	Negative label (0)
Dataset	241	458
Training set	169	320
Dev set	24	46
Test set	48	92

หมายเหตุ: Positive label (1) = Malignant และ Negative label (2) = Benign

มีโมเดลที่ใช้ในการพัฒนาดังนี้

KNN with manually tuning of k value model

ทำการปรับจูนค่า k พารามิเตอร์ด้วยการเทรน Training set แล้วนำไปทดสอบกับ Dev set เพื่อหา k ที่ดีที่สุด ได้ผลลัพธ์ดังนี้



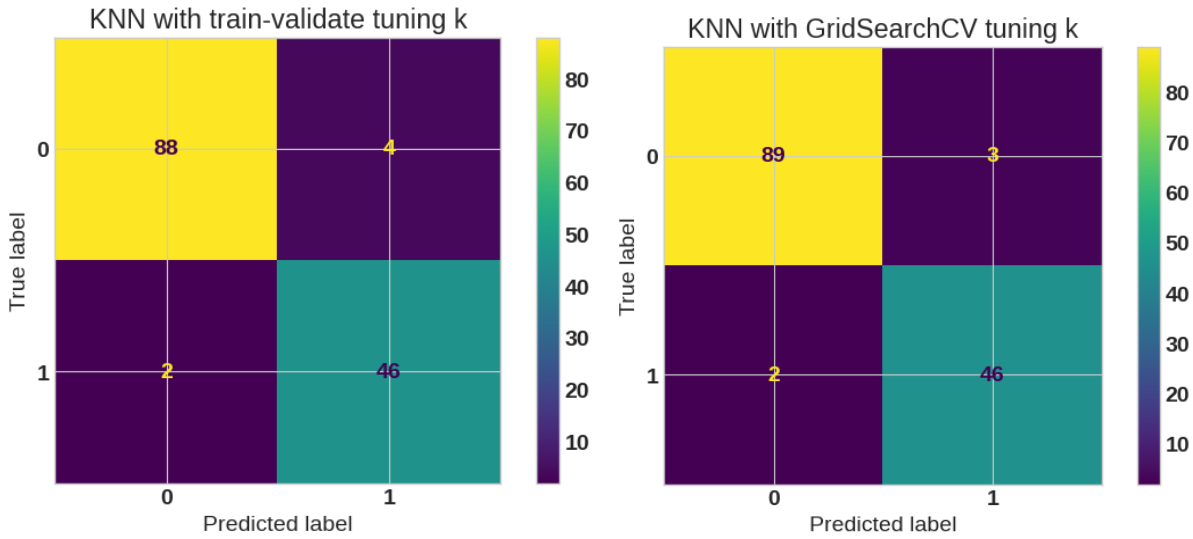
จากผลลัพธ์พบว่าค่า k ที่ดีที่สุดคือ k = 1, 3, 4 มี Accuracy ใน Dev set เท่ากัน เราจึงเลือก k = 1 เนื่องจากความประหยัดและความง่าย ดังนั้นจะนำค่า k นี้ไปใช้ในการเทรน Learning set (Training set + Dev set) แล้วนำไปทดสอบกับ Test set ต่อไป

KNN with gridsearch tuning of k value model

ใช้ library GridSearchCV จาก sklearn ในการเทรน Learning set เพื่อหาค่า k ที่ดีที่สุด ใช้ 5-fold cross validation และได้ค่า k ที่ดีที่สุดเท่ากับ 3 จากนั้นจึงนำค่า k นี้ทำไปทดสอบกับ Test set

ผลลัพธ์โมเดล

	Manually tuning of k value (k=1)	Gridsearch tuning of k value (k=3)
Test set	95.7	96.4



จากผลลัพธ์พบว่าโมเดลที่ปรับจูนพารามิเตอร์ด้วย GridsearchCV ได้ Accuracy สูงที่สุด แต่ก็ต่างจากโมเดลที่ปรับจูนพารามิเตอร์เองไม่มาก เพียงแค่ 0.7 เมื่อดูที่ confusion matrix ก็จะพบว่าต่างกันเพียงแค่การทายผิดว่าเป็น Malignant (False positive) ของโมเดลที่เราปรับจูนเองเพียงจุดเดียว นอกนั้นเหมือนกันหมด

สรุปผลได้ว่าโมเดลที่ปรับจูนพารามิเตอร์ด้วย GridsearchCV มีแนวโน้มที่จะทำได้ดีว่าโมเดลที่เราปรับจูนพารามิเตอร์เอง เนื่องจากโมเดลที่ปรับจูนพารามิเตอร์ด้วย GridsearchCV จะมีการใช้ cross-validation strategy คือการแบ่ง dataset ออกเป็นหลาย fold จากนั้นจึงเทรนโมเดลกับบาง fold และ evaluate โมเดลกับ fold ที่เหลือ จากนั้นจะสลับเทรนไปกับทุก fold อีกหลายครั้ง ทำให้โมเดลสามารถเรียนรู้จากชุดข้อมูลตัวอย่างใน learning set ได้ทุกส่วน จากมีแนวโน้มที่จะได้ประสิทธิภาพดีกว่าโมเดลที่เราปรับจูนพารามิเตอร์เองด้วยการแบ่ง learning set เป็น training set และ dev set ทำให้โมเดลไม่ได้เรียนรู้จากชุดข้อมูลตัวอย่างใน dev set เพื่อหาพารามิเตอร์ที่ดีที่สุดได้