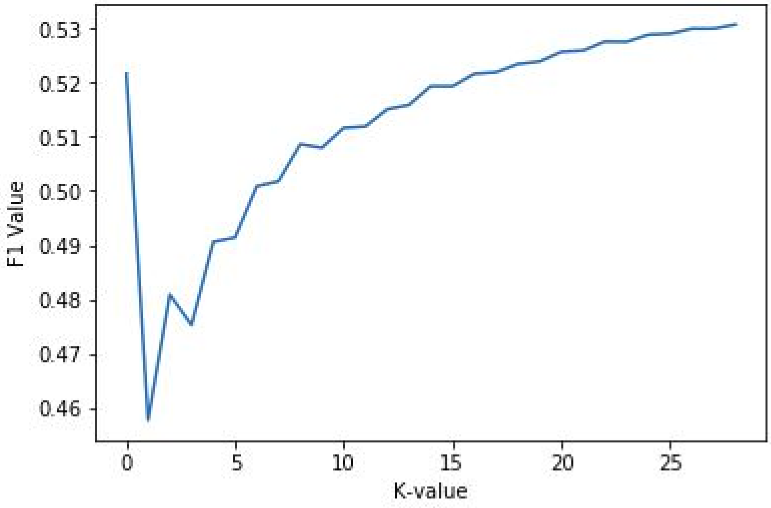
Lim, Dongjoon
260587899

1. Files submitted as zipfile on mycourses. See DS1.csv
2.
   (a) Performance measures are as below
   F1 score: 0.9547738693467337
   Accuracy: 0.9548872180451128
   Precision: 0.9547738693467337
   Recall: 0.9547738693467337
   (b) Parameters are as below
   w0: 25.664330858786784
   w1: [13.48509878093132, -7.99301198944706, -5.780340835989993, -3.442779729920341, -9.293118095905035, -4.028342669403219, 16.768342270361785, -22.194574854246987, -27.820990781388147, 8.214219012479413, -12.04395518639286, -12.132405121673251, 15.101046011468952, 12.371033704420272, -5.146378169241933, 12.14730620403903, 28.363843530991886, -6.126839937706375, -1.3602412970934328, -4.648949177698791]

3.
   (a) KNN performed worse than GDA. F1 value of GDA was well above 0.95 but F1 score for KNN was somewhere near 0.5. This is because GDA performs better for linearly separable model and KNN works well with low dimensions only. And yes there were certain K values that works slightly better than other K values, it is because when K is too small, overfitting can be occurred and when K is too large, small details of voting might be ignored. Thus it is important to find a sweet spot of K in between. In my case, the best K =29 has the largest F1 score. However, even the best K value did not give good F1 score so using KNN is not recommended.



(b) Performance measures are as below
F1 score: 0.5161290322580646
Accuracy: 0.53125
Precision: 0.5333333333333333
Recall: 0.5

4. See DS2.csv

DS2 and other files submitted on zipfile mycourses.

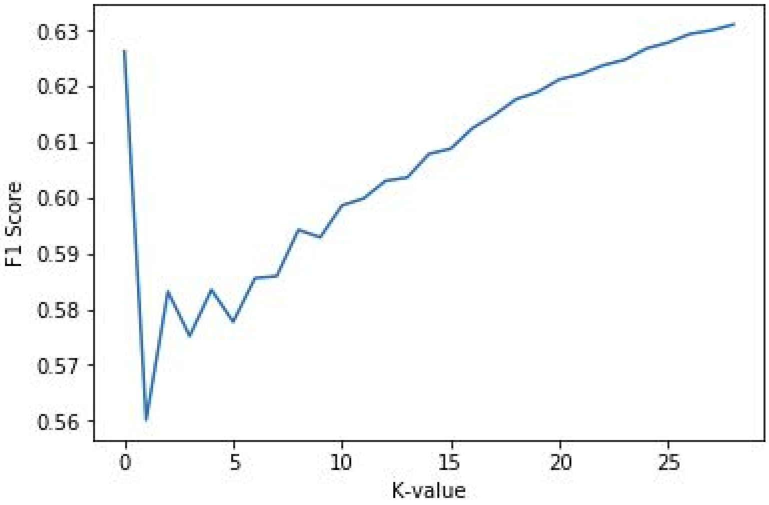
5. 1

(a) Performance measures are as below
F1 score: 0.5345268542199488
Accuracy: 0.545
Precision: 0.5471204188481675
Recall: 0.5225

(b) w0: -0.06401405481083025
w1: [-0.025346606777488734, 0.043182383061168866, -0.012434019476253014, 0.008213649615022661, 0.004914137583053327, -0.002412704657156914, 0.0016747489193872922, 0.004108037566534772, 0.007175541704667152, -0.03406544038308659, -0.013233958223824935, -0.008328029059978592, 0.07428917697616866, -0.0036420562809946766, -0.018500272834919756, 0.021692101472672837, 0.037000154351407744, -0.028897712346052123, 0.0034368016535646917, -0.0018154474685560529, 0.0]

5.2

This case, both GDA and KNN did not work well with poor F1 score. This mixed dataset is not linearly separable thus GDA does not work and KNN does not work due to curse of dimensionality. And yes again there were certain K values that works slightly better than other K values, it is because when K is too small, overfitting can be occurred and when K is too large, small details of voting might be ignored. Thus it is important to find a sweet spot of K in between. In my case, the best K = 29 has the largest F1 score. However, no matter what K values we choose, the F1 Score wasn't high enough so using KNN is not recommended.



5.3

Performance measures are as below
F1 score: 0.581453634085213
Accuracy: 0.5825
Precision: 0.5829145728643216
Recall: 0.58

6.

- For both DS1 and DS2, KNN did not perform well. This may be due to the high dimensionality of data (Having many features). For GDA, it worked well for DS1 but did not work well for DS2 where Dataset is a mixture of 3 gaussian. Because DS2 is a mixture of 3 gaussian, data become non linearly separable.