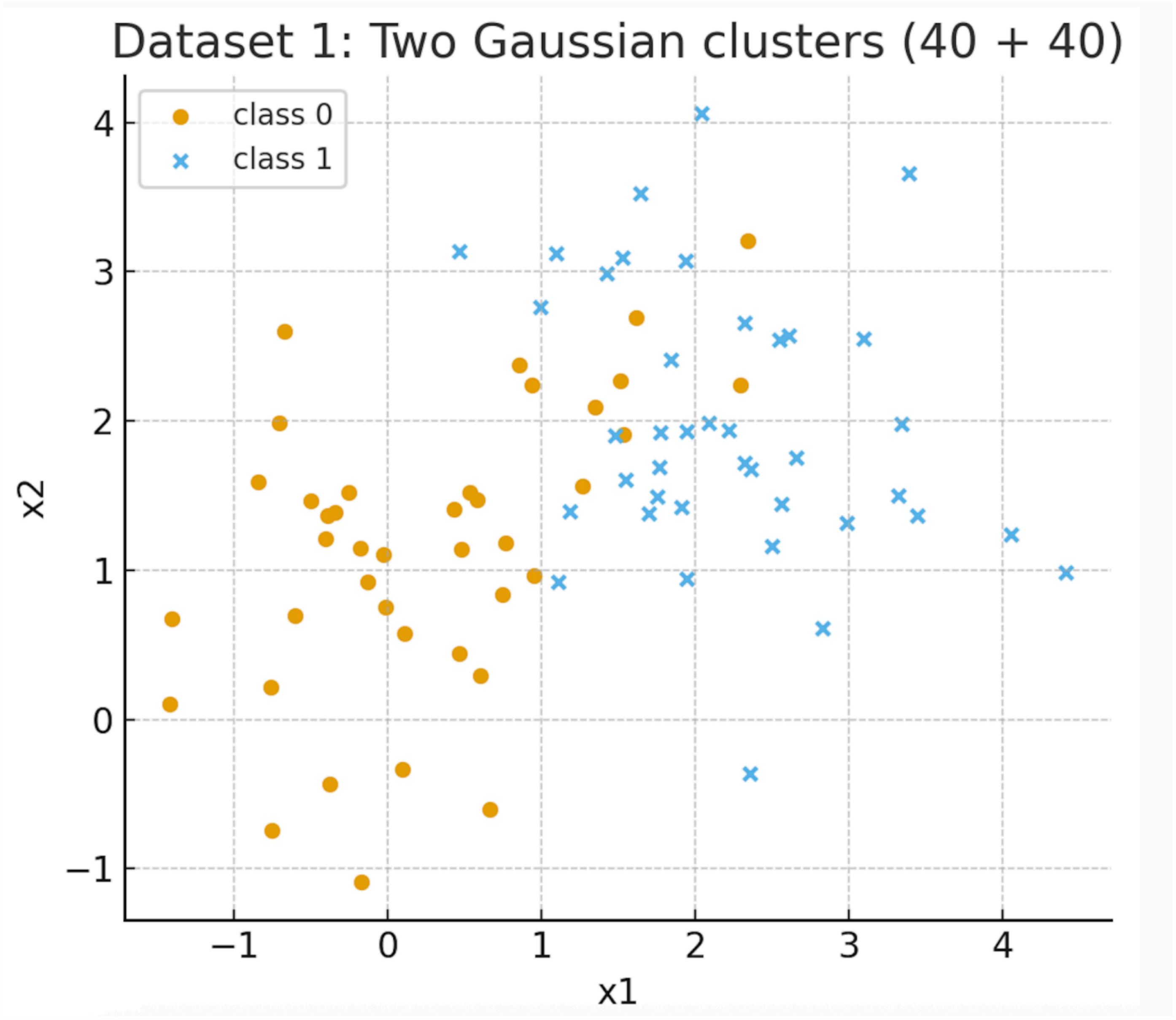
Part.B

1. Dateset1

Our Dataset1 contains 80 samples, in which there are 40 samples with `y=0`, and 40 samples with `y=1`. It was generated through Gaussian clusters, and then stored them on the disk in the Numpy `.npy` format.



feature1 **feature2** **label**

-0.1784637424485630	1.1442869652435100	0
0.7495904145465280	0.8352596500090170	0
0.9548472585837160	0.9631888877436280	0
-0.844555153601184	1.5936922615224200	0
0.7645416621333130	1.1807868743362400	0
-0.6060513743793620	0.6931615422735390	0
0.4679612044212450	0.44341087447755100	0
-0.38952443129654300	1.3673269879717600	0
1.3531627422439500	2.0961189147860400	0
2.2954452393098200	2.2375129860406800	0
1.6211827703905800	2.692674773323640	0
0.8584072176711300	2.375768390096750	0
-0.015044196592351700	0.7551524591177790	0
2.3445433562370900	3.2071710125134500	0
-0.028297874922390200	1.102774000298400	0
1.5146839634063200	2.268743328404230	0
1.2676392923050300	1.5662363932232000	0
-0.37367728045284100	-0.4342697042485070	0
-0.4992238039806130	1.4626238439138400	0
0.5358164393793030	1.5182463055524700	0
-0.12674976384362300	0.9226719075256420	0
0.9402238057078760	2.239144666991400	0
-0.17395691182432000	-1.086805174004270	0
-0.7622062694262290	0.21487420350894700	0
-0.6723670613360700	2.6036668507314100	0

0.0992803767020507	-0.33202666327168500	0
-0.40253307882607700	1.2077397337406500	0
-0.25383330388293200	1.5178275522556900	0
-0.3428805600759970	1.390229501683550	0
-0.7557852676418120	-0.7405821016333690	0
0.11182648751603800	0.575035566912016	0
0.6028934873446810	0.2969368325952510	0
0.5825533001319400	1.4728781626219000	0
-1.4033171544297800	0.6758468701916590	0
1.5367375776874200	1.9116828788417700	0
0.6632460452284320	-0.6051033525085560	0
0.43037655679380700	1.4070964194832200	0
-1.4207150621212900	0.10232869920551200	0
0.4821740550021290	1.1415694756345600	0
-0.7038135814186100	1.9874836356974500	0
2.365425000359470	1.6752555184148400	1
1.8442848471742100	2.412817635111410	1
1.6999646192922200	1.3818533601854700	1
1.9468627139564000	1.9282849049632900	1
1.6462100988309400	3.5265346386500900	1
2.6148847490220600	2.575954342809810	1
3.0964579405066000	2.548434755164580	1
2.5494869201246000	2.540780390152160	1
3.4458500897709200	1.3643472759038100	1
2.356956830283130	-0.3651905149064520	1

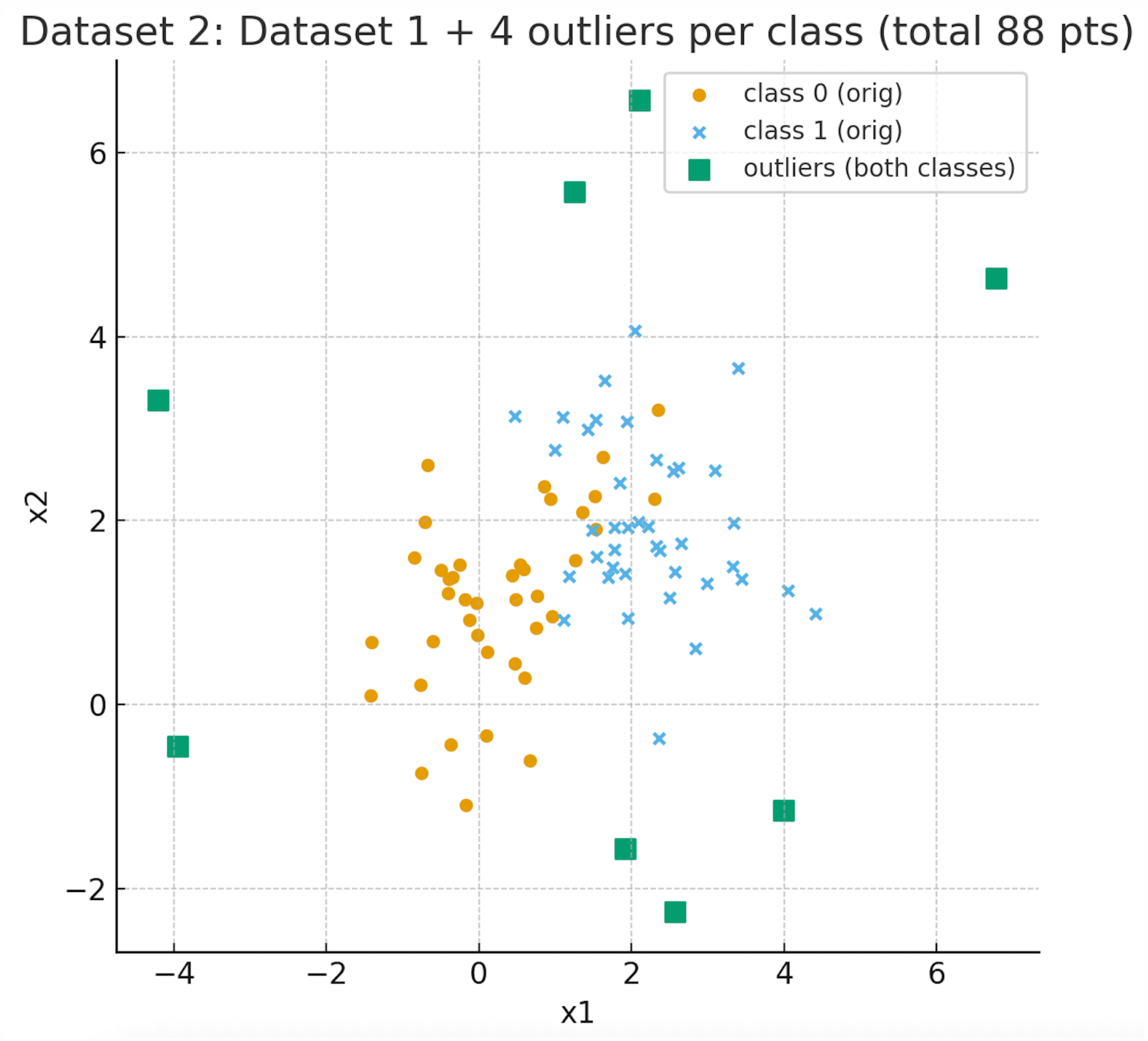
1.9130727540939500	1.4221185407926000	1
3.3957523372572100	3.660279521929390	1
2.5014266626223600	1.1603133298665500	1
2.3260686410432800	2.657272552181530	1
2.2203930908157500	1.9359468564766400	1
1.9412377657722000	3.0739252463348600	1
2.98653179881027	1.3197796015372600	1
1.5477093176871800	1.6046076514939800	1
1.4829371227873700	1.9019150833319300	1
1.5303247849070700	3.0977352544605200	1
1.7783417203563500	1.9223441017293200	1
1.1137436516352600	0.9224910573188100	1
3.326828784619190	1.5024890554777100	1
4.416965968383810	0.9847547432431060	1
4.055440519456520	1.239944551884510	1
3.342518840978810	1.9791958400798500	1
1.185219949178080	1.3963724267780000	1
2.0458295371032300	4.060021374084250	1
2.0910025593215600	1.9848953366663100	1
1.7519809742859800	1.4942032372672300	1
0.9931118402254740	2.764529763431160	1
1.7719505488366900	1.6894084885642800	1
1.9488438761179100	0.9393313362738170	1
1.0953588808686800	3.1223222437971200	1

1.4291320323571100	2.9866338433261900	1
2.5665764159532500	1.4468843716195500	1
2.6563813793534000	1.7530359538025100	1
2.325113718566920	1.7200141213939100	1
2.8341059548307400	0.611221069074144	1
0.46934924506403100	3.1364369028536300	1

2. Dataset2

We add four outliers for ach type (`y=0`and `y=1`) based on dataset1, repsecitvely. Hence, there are 88 samples on dataset2. The outliers are ploted on the pircutre below with green sqaure.

feature1	feature2	label
2.565142301299080	-2.2456412656282300	0
-3.951047159005270	-0.44764262073455400	0
1.2461230358784400	5.573061064777870	0
-4.209301764620870	3.313574892076170	0
1.9117748740417800	-1.5620015479615600	1
2.0987181302286900	6.5653514457518900	1
6.78426287967574	4.630375747339830	1
3.9908467663642600	-1.1429863003365900	1



3.Spliting Data

The dataset Dataset1 and Dataset2 were split using function `train_test_split` form `sklearn.model_selection`. Three critical parameters were used in this function:

- test_size=0.3, which means 30% samples in a dataset were used as its own testset.
- `stratify=type`, which is important because it guarantees that 30% of the samples are proportionally selected from both classes and used as the test set.
- random_state=42, which controls the randomness and ensures the reproductility of the traning process.

The split data of Dataset1

The table below is the test set of Dataset1:

feature1	feature2	label
0.6632460452284320	-0.6051033525085560	0
4.055440519456520	1.239944551884510	1
-0.3428805600759970	1.390229501683550	0
2.045829537103230	4.060021374084250	1
-0.1267497638436230	0.9226719075256420	0
-0.1739569118243200	-1.086805174004270	0
0.4693492450640310	3.1364369028536300	1
1.9468627139564	1.928284904963290	1
0.5825533001319400	1.4728781626219000	0
1.5477093176871800	1.6046076514939800	1
2.3260686410432800	2.657272552181530	1
2.8341059548307400	0.611221069074144	1
0.4821740550021290	1.1415694756345600	0
1.7783417203563500	1.9223441017293200	1
-0.0282978749223901	1.102774000298400	0
3.342518840978810	1.979195840079850	1
1.8442848471742100	2.412817635111410	1
-0.3895244312965430	1.3673269879717600	0
-0.4992238039806130	1.4626238439138400	0
3.445850089770920	1.3643472759038100	1
1.6211827703905800	2.692674773323640	0

-0.844555153601184	1.5936922615224200	0
-0.7622062694262290	0.2148742035089470	0
1.5303247849070700	3.0977352544605200	1

The table below is the training set of Dataset1:

feature1	feature2	label
2.98653179881027	1.3197796015372600	1
-0.402533078826077	1.207739733740650	0
1.6462100988309400	3.526534638650090	1
2.365425000359470	1.6752555184148400	1
1.113743651635260	0.9224910573188100	1
2.344543356237090	3.2071710125134500	0
2.0910025593215600	1.9848953366663100	1
-0.0150441965923516	0.7551524591177790	0
2.6148847490220600	2.575954342809810	1
2.2203930908157500	1.935946856476640	1
1.948843876117910	0.9393313362738170	1
2.325113718566920	1.7200141213939100	1
2.5494869201246000	2.540780390152160	1
-1.4033171544297800	0.6758468701916590	0
1.771950548836690	1.6894084885642800	1
1.2676392923050300	1.5662363932232000	0
2.2954452393098200	2.2375129860406800	0

0.4303765567938070	1.4070964194832200	0
1.4291320323571100	2.9866338433261900	1
0.8584072176711300	2.375768390096750	0
1.9412377657722000	3.0739252463348600	1
2.5014266626223600	1.160313329866550	1
0.4679612044212450	0.4434108744775510	0
0.7645416621333130	1.1807868743362400	0
1.9130727540939500	1.4221185407926000	1
0.111826487516038	0.575035566912016	0
1.5146839634063200	2.268743328404230	0
1.4829371227873700	1.9019150833319300	1
-0.3736772804528410	-0.4342697042485070	0
-0.1784637424485630	1.1442869652435100	0
0.0992803767020507	-0.332026663271685	0
1.185219949178080	1.3963724267780000	1
1.0953588808686800	3.1223222437971200	1
3.3957523372572100	3.660279521929390	1
1.6999646192922200	1.3818533601854700	1
2.5665764159532500	1.4468843716195500	1
-0.7038135814186100	1.9874836356974500	0
-1.4207150621212900	0.1023286992055120	0
0.993111840225474	2.764529763431160	1
0.5358164393793030	1.5182463055524700	0
3.096457940506600	2.548434755164580	1
3.326828784619190	1.5024890554777100	1

4.416965968383810	0.9847547432431060	1
0.9402238057078760	2.239144666991400	0
0.9548472585837160	0.9631888877436280	0
-0.7557852676418120	-0.7405821016333690	0
-0.6723670613360700	2.6036668507314100	0
1.7519809742859800	1.4942032372672300	1
1.3531627422439500	2.0961189147860400	0
-0.6060513743793620	0.6931615422735390	0
2.356956830283130	-0.3651905149064520	1
0.7495904145465280	0.8352596500090170	0
0.6028934873446810	0.2969368325952510	0
-0.2538333038829320	1.5178275522556900	0
1.5367375776874200	1.9116828788417700	0
2.6563813793534000	1.7530359538025100	1

2. The split data of Dataset2

The table below is the test set of Dataset2:

feature1	**feature2**	**label**
2.5494869201246000	2.540780390152160	1
0.9548472585837160	0.9631888877436280	0
-0.1739569118243200	-1.086805174004270	0
2.98653179881027	1.3197796015372600	1
3.445850089770920	1.3643472759038100	1

-0.3895244312965430	1.3673269879717600	0
2.565142301299080	-2.2456412656282300	0
1.9468627139564	1.928284904963290	1
2.045829537103230	4.060021374084250	1
2.0987181302286900	6.5653514457518900	1
1.8442848471742100	2.412817635111410	1
0.4303765567938070	1.4070964194832200	0
-0.7622062694262290	0.2148742035089470	0
-0.4992238039806130	1.4626238439138400	0
-0.3428805600759970	1.390229501683550	0
1.5303247849070700	3.0977352544605200	1
0.4821740550021290	1.1415694756345600	0
0.9402238057078760	2.239144666991400	0
3.096457940506600	2.548434755164580	1
4.055440519456520	1.239944551884510	1
2.3260686410432800	2.657272552181530	1
1.5367375776874200	1.9116828788417700	0
-0.1267497638436230	0.9226719075256420	0
3.9908467663642600	-1.1429863003365900	1
-0.0282978749223901	1.102774000298400	0
1.6211827703905800	2.692674773323640	0
3.342518840978810	1.979195840079850	1

The table below is the traning set of Dataset2:

feature1	feature2	label

1.5146839634063200	2.268743328404230	0
2.2954452393098200	2.2375129860406800	0
0.0992803767020507	-0.332026663271685	0
-3.951047159005270	-0.44764262073455400	0
-4.209301764620870	3.313574892076170	0
2.365425000359470	1.6752555184148400	1
0.4693492450640310	3.1364369028536300	1
-0.7557852676418120	-0.7405821016333690	0
4.416965968383810	0.9847547432431060	1
2.325113718566920	1.7200141213939100	1
2.356956830283130	-0.3651905149064520	1
0.5825533001319400	1.4728781626219000	0
1.9130727540939500	1.4221185407926000	1
-0.844555153601184	1.5936922615224200	0
1.2676392923050300	1.5662363932232000	0
2.6148847490220600	2.575954342809810	1
-1.4207150621212900	0.1023286992055120	0
0.5358164393793030	1.5182463055524700	0
1.113743651635260	0.9224910573188100	1
1.4829371227873700	1.9019150833319300	1
0.7495904145465280	0.8352596500090170	0
6.78426287967574	4.630375747339830	1
-0.6060513743793620	0.6931615422735390	0
0.6028934873446810	0.2969368325952510	0
2.8341059548307400	0.611221069074144	1

3.326828784619190	1.5024890554777100	1
-0.402533078826077	1.207739733740650	0
1.6462100988309400	3.526534638650090	1
2.2203930908157500	1.935946856476640	1
1.6999646192922200	1.3818533601854700	1
-0.2538333038829320	1.5178275522556900	0
2.344543356237090	3.2071710125134500	0
1.948843876117910	0.9393313362738170	1
-0.6723670613360700	2.6036668507314100	0
2.5665764159532500	1.4468843716195500	1
3.3957523372572100	3.660279521929390	1
0.993111840225474	2.764529763431160	1
0.8584072176711300	2.375768390096750	0
0.6632460452284320	-0.6051033525085560	0
1.185219949178080	1.3963724267780000	1
1.7783417203563500	1.9223441017293200	1
2.5014266626223600	1.160313329866550	1
1.0953588808686800	3.1223222437971200	1
1.2461230358784400	5.573061064777870	0
0.4679612044212450	0.4434108744775510	0
2.6563813793534000	1.7530359538025100	1
-0.0150441965923516	0.7551524591177790	0
1.7519809742859800	1.4942032372672300	1
2.0910025593215600	1.9848953366663100	1
-0.1784637424485630	1.1442869652435100	0

1.3531627422439500	2.0961189147860400	0
0.7645416621333130	1.1807868743362400	0
0.111826487516038	0.575035566912016	0
-0.7038135814186100	1.9874836356974500	0
1.771950548836690	1.6894084885642800	1
-0.3736772804528410	-0.4342697042485070	0
-1.4033171544297800	0.6758468701916590	0
1.9117748740417800	-1.5620015479615600	1
1.9412377657722000	3.0739252463348600	1
1.5477093176871800	1.6046076514939800	1
1.4291320323571100	2.9866338433261900	1

The four models:k-NN, Naive Bayes, Decision Trees, and Random Forests

1. Naive Bayes

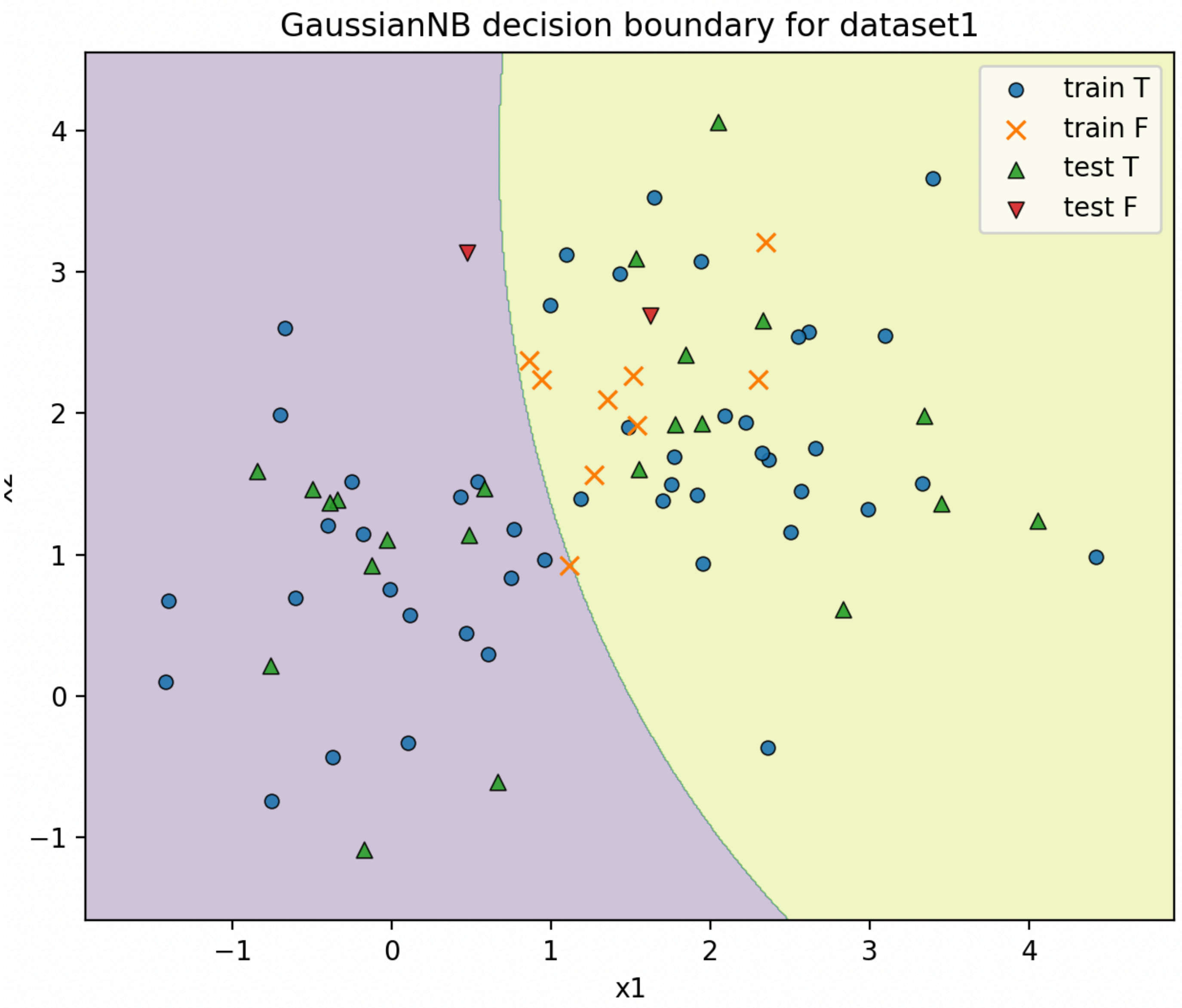
The Accuracy:

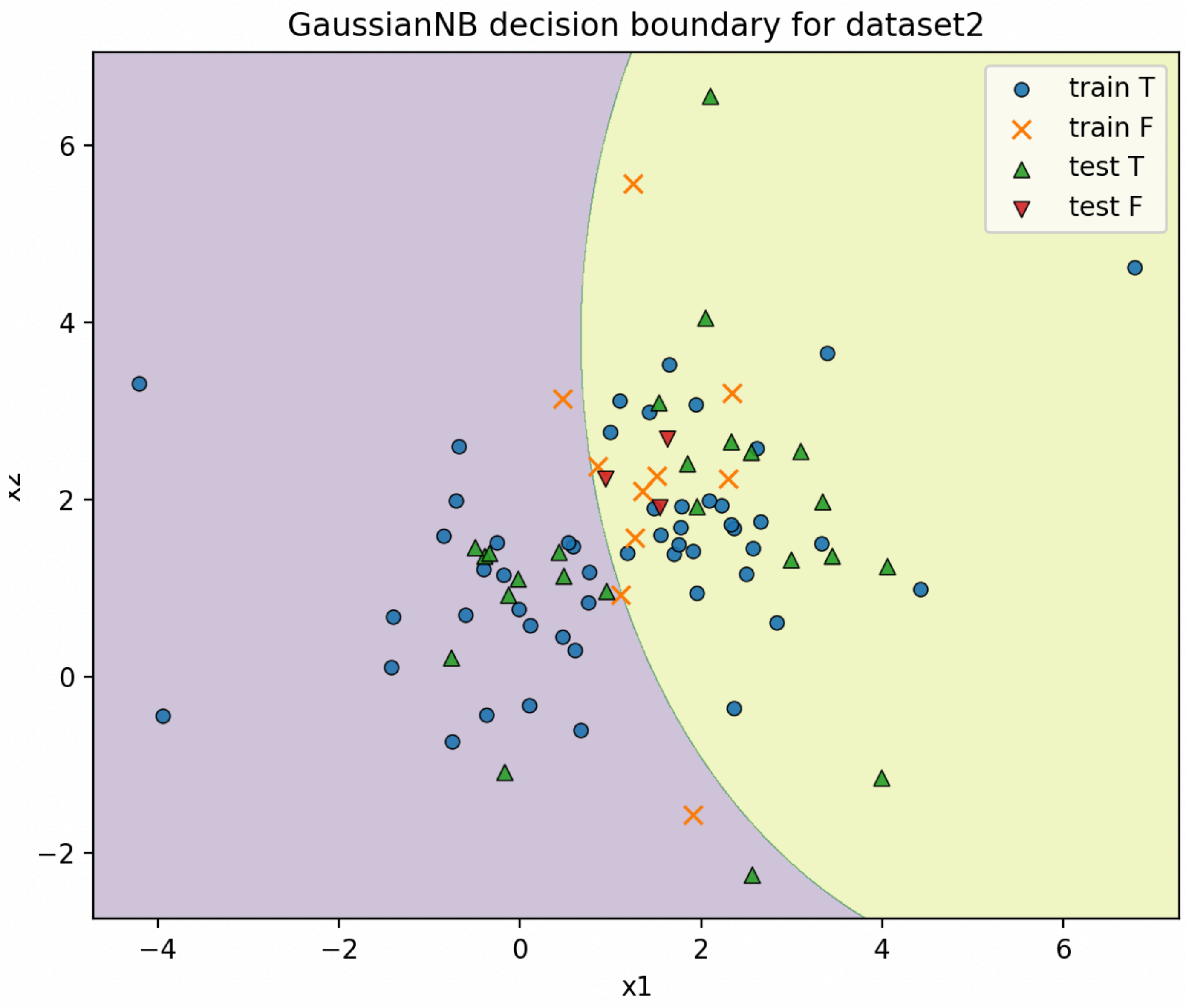
Dataset	Testing set	Training set
Dataset1	91.6%	85.7%
Dataset2	92.6%	83.6%

The Confusion Matrix:

Dataset	**Testing set**	**Training set**
Dataset1	$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$	$\begin{bmatrix} 22 & 6 \\ 2 & 26 \end{bmatrix}$
Dataset2	$\begin{bmatrix} 12 & 2 \\ 0 & 13 \end{bmatrix}$	$\begin{bmatrix} 24 & 6 \\ 4 & 27 \end{bmatrix}$

Visualization of the results:





2. Decision Tree

The Accuracy:

Dataset	Testing set	Training set
Dataset1	91.6%	100%
Dataset2	77.8%	100%

The Confusion Matrix:

Dataset	**Testing set**	**Training set**
Dataset1	$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$	$\begin{bmatrix} 28 & 0 \\ 0 & 28 \end{bmatrix}$
Dataset2	$\begin{bmatrix} 10 & 4 \\ 2 & 11 \end{bmatrix}$	$\begin{bmatrix} 30 & 0 \\ 0 & 31 \end{bmatrix}$

3. Random Forest

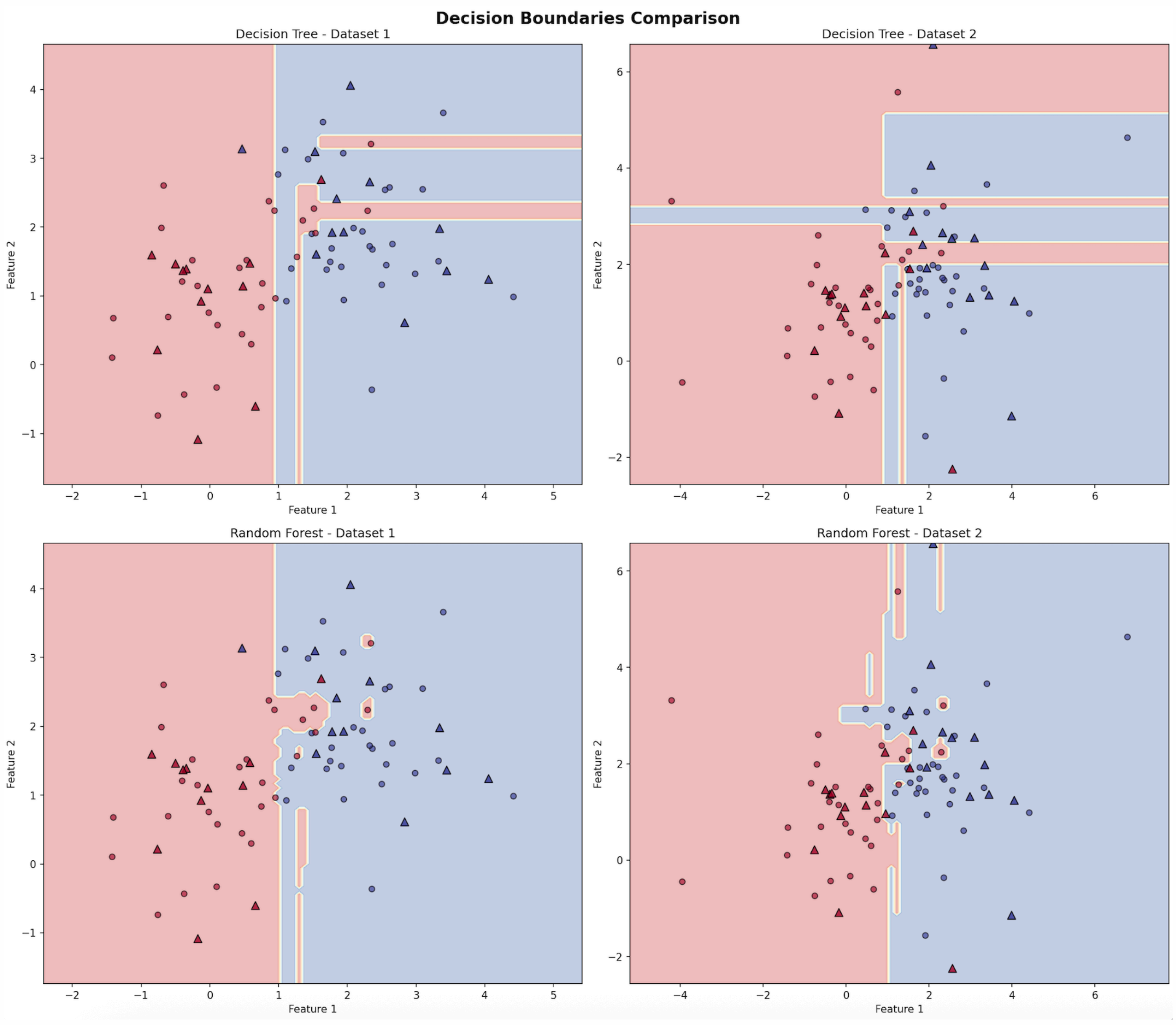
The Accuracy:

Dataset	**Testing set**	**Training set**
Dataset1	91.6%	100%
Dataset2	85.2%	100%

The Confusion Matrix:

Dataset	**Testing set**	**Training set**
Dataset1	$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$	$\begin{bmatrix} 28 & 0 \\ 0 & 28 \end{bmatrix}$
Dataset2	$\begin{bmatrix} 10 & 4 \\ 0 & 13 \end{bmatrix}$	$\begin{bmatrix} 30 & 0 \\ 0 & 31 \end{bmatrix}$

The visualization of results obtained from Decsion Tree and Random Forest:



4. K-NN

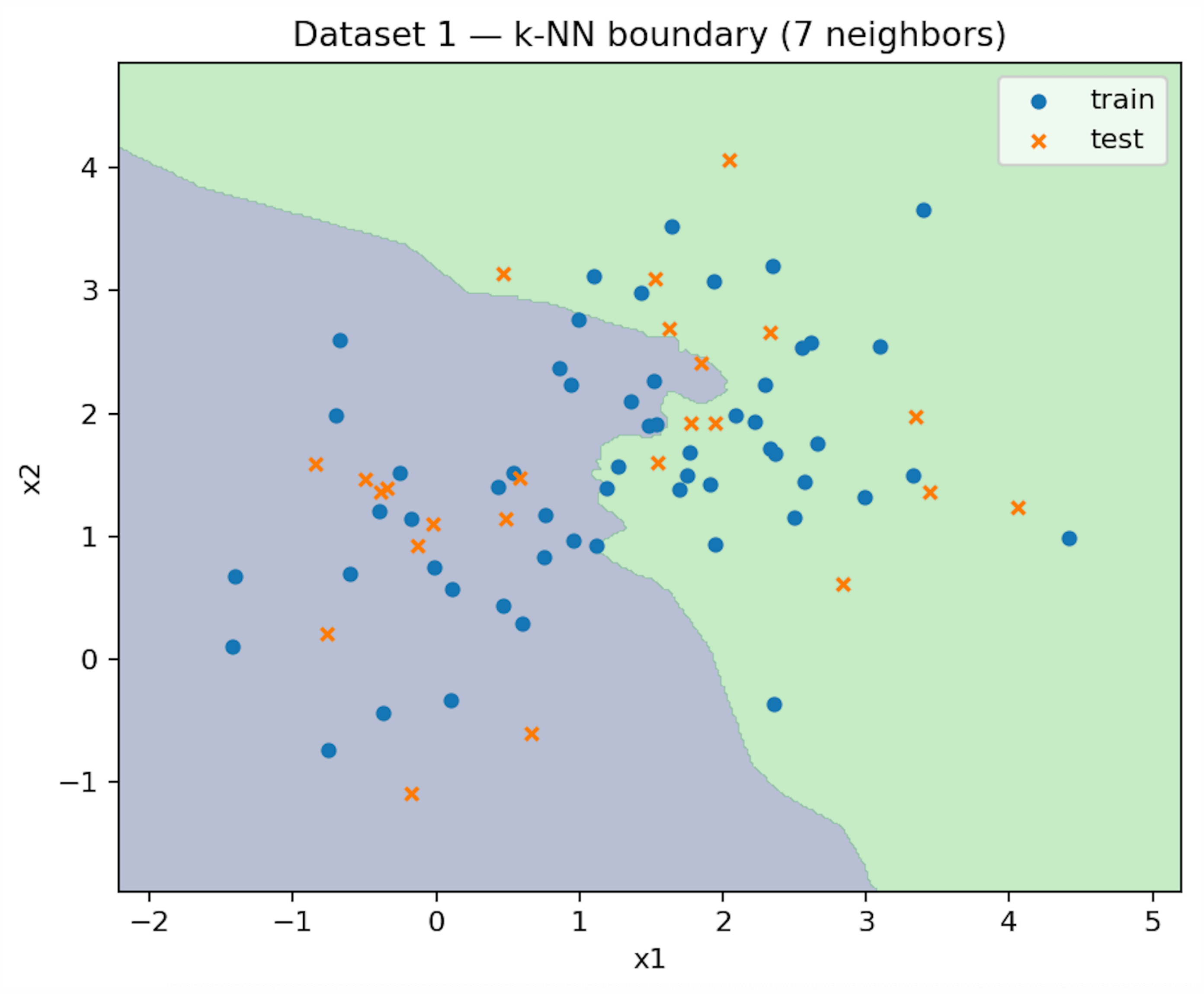
The Accuracy:

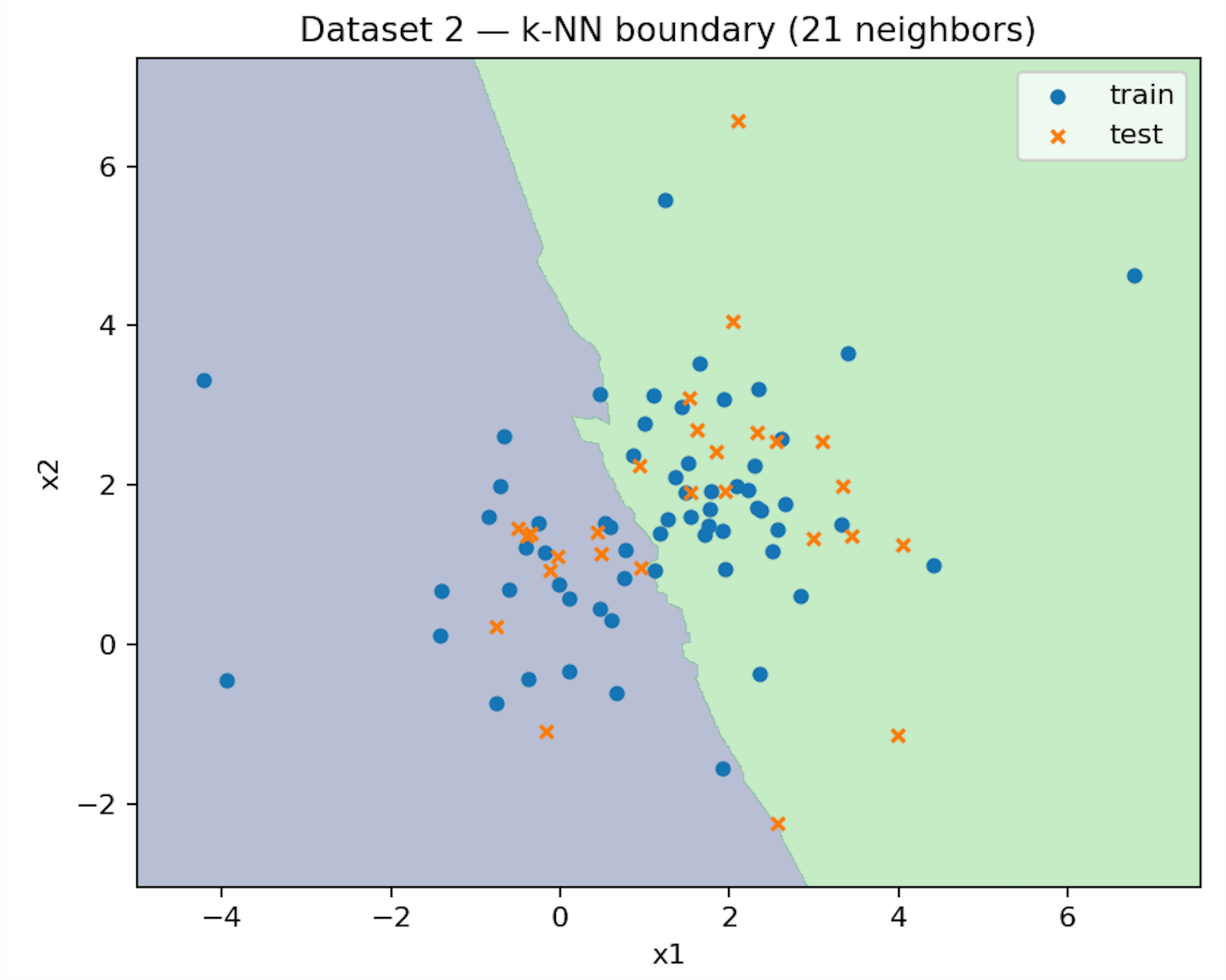
Dataset	Testing set	Training set
Dataset1	91.7%	91%
Dataset2	85.2%	85.2%

The Confusion Matrix:

Dataset	Testing set	Training set
Dataset1	$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$	$\begin{bmatrix} 25 & 3 \\ 2 & 26 \end{bmatrix}$
Dataset2	$\begin{bmatrix} 10 & 4 \\ 0 & 13 \end{bmatrix}$	$\begin{bmatrix} 23 & 7 \\ 2 & 29 \end{bmatrix}$

The visualization of results obtained from Decsion Tree and Random Forest:





Discussion about the results

We discussed the results from the **discrimination** and **generalization** perspective.

In general, Naive Bayes shows best performance acorss the datasets among the models. By contrast, Decision Tree performs the worst.

For this dataset, Naive Bayes model achieves the best performance on both perspectives, as they has highest test-set accurcay, with 91.6% and 92.6%, respectively. As both classes have equal numbers of samples, it is enough to use accuracy alone to evaluate their discrimination ability. This model's generalization ability is also excellent on the datasets. We can see this from the accuracy differences between testing sets and training sets: -5.9% and -9.0%.

The ability to discriminate of Random Forest is also strong. For Dataset1, its accuracy is 91.7%, the same as that of Naive Bayes. However, it only obtains 85.2% accuracy on Dataset2. In addition, this model slightly suffers from overfitting on Dataset2.

The discrimination ability of k-NN is almost smiliar to that of Random forest, but the other performance is better than Random forest's, because a reasonable accuracy differences.

Fininally, there is a large accuracy gap of 22.2% between the traning set and testing set on Dataset2, and the accuracy on training set is 100%. It is obvious that this model overfits Dataset2. In addtion, its discrimination performance on Dataset2 is also the poorest, because of a test accuracy with only 77.8%.

Similarities and dissimilarities

1. Similarities

These four models mainly have two similarites:

1. Supervised Learning: All of them are supervised learning, which means they learn from labeled data.
2. Classification Task: Their only goal is to perform classification, which means assigning new, unseen data points to one or more predefined classes.

2. Disimilarities

The main differences among them are their principles. Hence, this cause other important differcences: **Data-Model fit, Decision Boundary, Genralization perfomance and Cost.**

- **Data-Model fit**: What kind of data is suitable for this model?
  - k-NN: Thise model is suitable for data where distances in the feature space are meaningful.

    - Naive Bayes: This model expects the features to be as independent as possible.
    - Decesion Tree: This model expects the data to be divided into pure subsets through features, and does not require independence or linear relationships among them.
    - Random Forest: This model expects the data to be separable through different combinations of features, which provide complementary information.
- **Decision Boundary**:
    - k-NN: Its Decision Boundary is highly nonlinear and locally adaptive and entirely determined by the positions of the training samples.
    - Naive Bayes: The decision boundary is usually smooth and approximately linear or quadratic.
    - Decesion Tree: The decision boundary of a decision tree consists of horizontal and vertical lines parallel to the coordinate axes, forming a staircase-like pattern.
    - Random Forest: The decision boundary can be highly complex locally, but smoother and more generalizable globally than that of a single decision tree.
- **Cost on Traning and Predition**: Aussme that d is the demention of the features, n is the number of samples and T is the number of trees in the forest.
    -

Model	**Training Cost**	**Prediction Cost**
k-NN	**Low** — $O(1)$	**High** — $O(n \cdot d)$
Naive Bayes	**Low** — $O(n \cdot d)$	**Low** — $O(d)$
Decision Tree	**Medium** — $O(n \cdot d \log n)$	**Low** — $O(\text{tree depth})$
Random Forest	**High** — $O(T \cdot n \cdot d \log n)$	**Medium** — $O(T \cdot \text{tree depth})$

- **Genralization perfomance** :
    -

Model	Bias–Variance Trade-off	Sensitivity to Noise/Outliers	Typical Generalization
k-NN	**Low bias, high variance**	**Sensitive**: strongly affected by noisy neighbors, outliers, and feature scaling or distance metric	**Moderate**: strong on low-dimensional data with clear local structure; degrades in high-dimensional settings
Naive Bayes	**High bias, low variance**	**Relatively stable**	**Moderately favorable**: often excellent on high-dimensional sparse data (text); drops when features are strongly correlated
Decision Tree	**Low bias, high variance**	**Sensitive**	**Moderate** when well-regularized; but prone to overfitting without it
Random Forest	**Variance greatly reduced**; bias similar to or slightly higher than a single tree	**Very Stable**: relatively insensitive to noise or outliers	**High**: typically strong generalization performance