

# Lab10

## GDC Data Lab

By  
Siva Sai Krishna Marthy  
Jianan gao  
Xin Wang

<https://github.com/appleshine77/GDC-data-lab.git>

# Part1: Data download, integration and preprocess

**Disease Type**

- Adenomas and Adenocarcinomas 12,492
- Epithelial Neoplasms, NOS 3,202
- Ductal and Lobular Neoplasms 3,025
- Squamous Cell Neoplasms 2,635
- Gliomas 2,034
- Cystic, Mucinous and Serous Neoplasms 1,658
- Neuroblastoma 1,127
- Nevi and Melanomas 1,121
- Acute Myeloid Leukemia 988
- Transitional Cell Papillomas and Carcino... 813
- High-Risk Wilms Tumor 652
- Lymphoid Neoplasm Diffuse Large B-cell... 534
- Osteosarcoma 381
- Mesothelial Neoplasms 271
- Not Reported 255
- Acinar Cell Neoplasms 247
- Complex Mixed and Stromal Neoplasms 230
- Myeloid Leukemias 200
- Paragangliomas and Glomus Tumors 197

**Data Category**

- Transcriptome Profiling 11,486

**Data Type**

- Isoform Expression Quantification 11,486
- miRNA Expression Quantification 11,486

**Experimental Strategy**

- miRNA-Seq 11,486

**Workflow Type**

- BCGSC miRNA Profiling 11,486

**Data Format**

- TXT 11,486

**Platform**

No data for this field

**Access**

- open 11,486

[Add All Files to Cart](#) [Manifest](#) [View 10,601 Cases in Exploration](#) [View Images](#) [Browse](#)

Files (11,486)
Cases (10,601)

Primary Site
Project
Data Category
Data Type
Data Form

Show More

Showing 1 - 20 of 11,486 files

File UUID	Access	File Name	Cases	Project	Filter Columns
<a href="#">0e22d215-92c2-435c-ad89-ee344ef88635</a>	<a href="#">open</a>	<a href="#">9a2956fd-6985-4ed9-96a5-e04a3e</a>	<a href="#">1</a>	<a href="#">TCGA-LUSC</a>	<a href="#">Restore Defaults</a>
<a href="#">50efcf8f-1e01-4f71-b9e4-f8491eb3c736</a>	<a href="#">open</a>	<a href="#">95532a5a-df6a-438a-a26d-7c4a1c2</a>	<a href="#">1</a>	<a href="#">TCGA-LUSC</a>	<a href="#">File UUID</a>
<a href="#">bb2acdaa-3a26-421a-a9e9-4ea3336a467b</a>	<a href="#">open</a>	<a href="#">8051f857-9875-427f-a757-4091db3</a>	<a href="#">1</a>	<a href="#">TCGA-LUSC</a>	<a href="#">Access</a>
<a href="#">be8861c3-48b9-4cb7-a9a5-d4e18449d3ce</a>	<a href="#">open</a>	<a href="#">c591a.mirbase21.mirnas.quantification.txt</a>	<a href="#">1</a>	<a href="#">TCGA-LUSC</a>	<a href="#">File Name</a>
<a href="#">7ae1a03b-cca0-4ad6-92b8-096e24</a>	<a href="#">open</a>	<a href="#">7ae1a03b-cca0-4ad6-92b8-096e24</a>	<a href="#">1</a>	<a href="#">TCGA-LUSC</a>	<a href="#">Cases</a>
					<a href="#">Project</a>
					<a href="#">Data Category</a>
					<a href="#">Data Format</a>
					<a href="#">Size</a>
					<a href="#">Annotations</a>

Disease Type selects all in case page, and in file page, only select miRNA, there are totally 11,486 files and 10,601 cases;

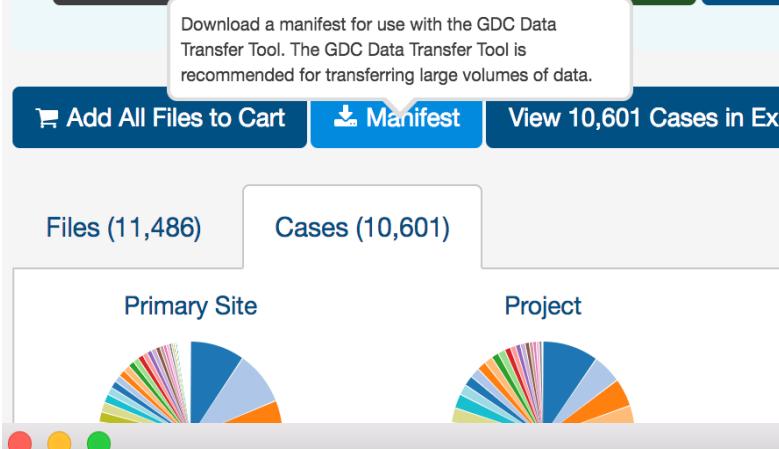
# Part1: Data download, integration and preprocess

Download a manifest for use with the GDC Data Transfer Tool. The GDC Data Transfer Tool is recommended for transferring large volumes of data.

Add All Files to Cart   Manifest   View 10,601 Cases in Export

Files (11,486)   Cases (10,601)

Primary Site   Project



id	filename	md5	size	state
0e22d215-92c2-435c-ad89-ee344ef88635	9a2956fd-6985-4ed9-96a5-e04a3ed3acbb.mirbase21.mirnas.quantification.txt	95532a5a-df6a-438a-a26d-7c4a1c241c9f.mirbase21.mirnas.quantification.txt	8051f857-9875-427f-a757-4091db3c591a.mirbase21.mirnas.quantification.txt	7ae1a03b-cca0-4ad6-92b8-096e247c1b50.mirbase21.mirnas.quantification.txt
50efcf8f-1e01-4f71-b9e4-f8491eb3c736	37be9876-e459-48e9-b9ba-b47d02c17ae7.mirbase21.mirnas.quantification.txt	69c2dd2f-9df4-4f79-9d18-577ad2d53dfd.mirbase21.mirnas.quantification.txt	7df584d3-5b11-4987-8227-6d809e8305af.mirbase21.mirnas.quantification.txt	36ff000c-51b2-4e3d-9300-4c536d0ae141.mirbase21.mirnas.quantification.txt
bb2acdaa-3a26-421a-a9e9-4ea3336a467b	cb606a2b-3c67-41c8-866a-2422e7ac369f.mirbase21.mirnas.quantification.txt	1f39d8bd-9c1f-4dcf-9de0-62e3966b014f.mirbase21.mirnas.quantification.txt	f7ca0295-d9da-4a0d-8923-8a2053aedfe3.mirbase21.mirnas.quantification.txt	d3394983-097e-45dd-b449-0dded9446dc.d.mirbase21.mirnas.quantification.txt
be8861c3-48b9-4cb7-a9a5-d4e18449d3ce	ad29232f-e5e5-4206-8ab3-cf45ac0c49e1.mirbase21.mirnas.quantification.txt	b1475bd4-e50b-41bf-a2dd-19bcd035805b.mirbase21.mirnas.quantification.txt	944e5ba8-7c95-4b21-81eb-c02b4eb0bbed.mirbase21.mirnas.quantification.txt	064d7af1-f631-42aa-bfa6-23afabdee202.mirbase21.mirnas.quantification.txt
0435e383-6ec2-46c4-ba15-713a2447b0d1	48053f52-5a54-417c-abe6-190e236a47df.mirbase21.mirnas.quantification.txt	48053f52-5a54-417c-abe6-190e236a47df.mirbase21.mirnas.quantification.txt	72bb1711-cec9-4de1-ad36-80b806dbcddde.mirbase21.mirnas.quantification.txt	50271141-0251-4415-9201-3cd1-8007-50b
f7e016d-2d0b-40d1-a330-09ae95684813	50309	released		
07fa22d0-99b1-4364-9b93-d937005a2416	50362	released		
d6e60c88-6acc-45a0-abe8-e8bcfbccf369	50206	released		
2c5eacad-0a4f-43e6-ad8f-dae2ffe4b0c6	50277	released		
5f06f29b-ddc7-43ae-a2d0-7df4349f1507	50426	released		
94a5d869-a95f-4dca-9f5a-59aaaeb39670	50307	released		
314186da-49ad-4db7-b3c2-c1e0b262f4f8	50150	released		
ece4bd52-0c7b-4cb8-ab6c-fbe10f97eb36	50392	released		
c7cf5c61-54b6-41ea-8c8e-294211f2b7fa	50297	released		
5984e36f-e65b-4095-a1f9-c848bc0fe08a	6343156b82a3deaa4ef11098d7e2d1c5	50231	released	
ac3a3cb1-b6ce-41a8-af04-132012e48df3	65eed838d29779714f8c2ce8575225b7	50230	released	
ae4cd42a-c88a-4e9d-8a19-f93a9ade179a	cbcc855aa63622a6bad762ac2a66a2ca	50267	released	
54057a72-d718-4ba7-9e04-68036c16a79f	e41ac4c76f0c8c490f01d8a0c4459152	50424	released	
fe335358-a468-49b1-89ec-31e1b5e084bd	1618d7a51dcfc8cd3c4af088d544bb	50207	released	
50271141-0251-4415-9201-3cd1-8007-50b	ba3a546250b44cdbc9e4210061eae37e	50321	released	
	162d0e4f1bf2092cb32ea9dedda3a22b	50340	released	
	f1da8f5fbb352d74c6f6312673f5a48	50103	released	
	fde3b7bc7791af6d70635884cbc01a0	50126	released	
	32c5d6dce8fba6d739ae158fd5c9e4d	50276	released	
	7d6a4f51faed7dc9380b876ebb5a90d8	50250	released	

gdc\_manifest.2018-10-21.txt

Manifest download

# Part1: Data download, integration and preprocess

```
[XINs-MacBook-Pro:miRNA xin$ python -3 ../../src/check.py
[...]
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:16: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
    from . import multiarray
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:34: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
    from . import umath
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/_internal.py:204: DeprecationWarning: Overriding __e
q__ blocks inheritance of __hash__ in 3.x

/Users/xin/anaconda/lib/python2.7/site-packages/pandas/core/reshape/util.py:76: DeprecationWarning: reduce() no
t supported in 3.x; use functools.reduce()
    return reduce(_compose2, funcs)
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/pytables.py:1488: DeprecationWarning: Overriding __eq
__ blocks inheritance of __hash__ in 3.x
    class IndexCol(StringMixin):
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/pytables.py:1790: DeprecationWarning: Overriding __eq
__ blocks inheritance of __hash__ in 3.x
    class DataCol(IndexCol):
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/stata.py:720: DeprecationWarning: Overriding __eq__ b
locks inheritance of __hash__ in 3.x
    class StataMissingValue(StringMixin):
[2018-10-21 11:58:03,798 - GDC - INFO] ===start checking===
[2018-10-21 11:58:46,772 - GDC - INFO] successful downloads
[2018-10-21 11:58:46,773 - GDC - INFO] ===check finished==
```

Check the successful download by check.py

# Part1: Data download, integration and preprocess

<https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>

Links to the binary distributions for supported platforms

-  [gdc-client\\_v1.3.0\\_Windows\\_x64.zip](#)
-  [gdc-client\\_v1.3.0\\_Ubuntu14.04\\_x64.zip](#)
-  [gdc-client\\_v1.3.0 OSX\\_x64.zip](#)

```
XINs-MacBook-Pro:miRNA xin$ ../../gdc-client download -m ../gdc_manifest.2018-10-21.txt
```

```
[  
  
100% [########################################] Time: 0:00:04 0.24 B/s  
100% [########################################] Time: 0:00:04 0.25 B/s  
100% [########################################] Time: 0:00:04 0.24 B/s  
100% [########################################] Time: 0:00:04 0.23 B/s  
100% [########################################] Time: 0:00:04 0.22 B/s  
100% [########################################] Time: 0:00:04 0.24 B/s  
100% [########################################] Time: 0:00:03 0.26 B/s  
100% [########################################] Time: 0:00:03 0.26 B/s  
  
100% [########################################] Time: 0:00:01 39.65 kB/s  
100% [########################################] Time: 0:00:00 58.05 kB/s  
100% [########################################] Time: 0:00:00 67.05 kB/s  
100% [########################################] Time: 0:00:00 62.62 kB/s  
100% [########################################] Time: 0:00:00 63.52 kB/s  
100% [########################################] Time: 0:00:00 72.58 kB/s  
100% [########################################] Time: 0:00:00 72.68 kB/s  
100% [########################################] Time: 0:00:00 84.89 kB/s  
100% [########################################] Time: 0:00:00 73.23 kB/s  
100% [########################################] Time: 0:00:00 70.54 kB/s  
  
Successfully downloaded: 11486
```

Download GDC-data-transfer tool and download gdc\_manifest file with the tool

# Part1: Data download, integration and preprocess

The screenshot shows a user interface for managing data files. On the left, there is a sidebar with a search bar labeled 'Filter Columns' and a 'Restore Defaults' button. Below these are several checkboxes for selecting items: 'Select', 'Cart', 'Case UUID', 'Case ID', 'Project', 'Primary Site', 'Gender', 'Files', 'Data Categories', 'Annotations', 'Slides', and 'Program'. Most of these checkboxes are checked. On the right, there is a table titled 'Available Files per Data Category' with columns: Seq, Exp, SNV, CNV, Meth, and Cl. The table lists several rows of file counts, such as 39, 2, 5, 0, 4, 2; 57, 4, 5, 16, 4, 1; etc. At the top right of the interface, there are buttons for 'Export All', 'Biospecimen', and 'JSON' (which is highlighted).

	Seq	Exp	SNV	CNV	Meth	Cl
1	39	2	5	0	4	2
2	57	4	5	16	4	1
3	38	2	5	0	4	2
4	65	6	5	24	4	1
5	53	4	5	16	4	1
6	76	7	10	24	6	2
7	53	4	5	16	4	1
8	55	4	5	16	4	2
9	39	3	5	0	4	1
10	38	2	5	0	4	1
11	54	4	5	16	4	2

The screenshot shows a JSON file named 'files.2018-10-21.json' being viewed in a file editor. The file contains two entries, each representing a transcriptome profiling file. The first entry has a file name of '544f8d81-daea-4571-bac2-c46cdada8405.mirbase21.mirnas.quantification.txt', a data format of 'TXT', access of 'open', a file ID of '52792628-b48a-4068-878c-960d5f6ebcd', a data category of 'Transcriptome Profiling', a file size of 50498, and a case ID of '5a2f8140-8f90-4e94-b703-5fa5aa96be7b'. The second entry has a file name of '63ddf182-d8df-4a46-9231-6090272d2afa.mirbase21.mirnas.quantification.txt', a data format of 'TXT', access of 'open', a file ID of 'eb73f062-1fe7-484e-ab22-856d2fd936ad', a data category of 'Transcriptome Profiling', a file size of 50154, and a case ID of 'beb0025-74e2-451b-93b3-86f82df43573'.

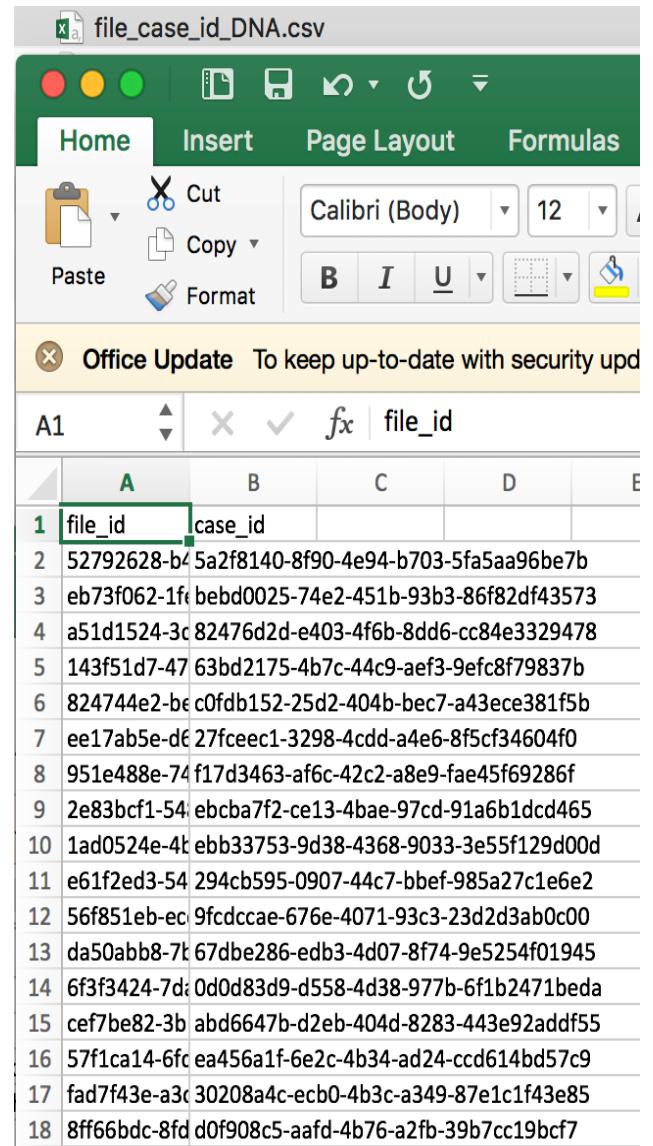
```
[{"file_name": "544f8d81-daea-4571-bac2-c46cdada8405.mirbase21.mirnas.quantification.txt", "data_format": "TXT", "access": "open", "file_id": "52792628-b48a-4068-878c-960d5f6ebcd", "data_category": "Transcriptome Profiling", "file_size": 50498, "cases": [{"project": {"project_id": "TCGA-LUAD"}, "case_id": "5a2f8140-8f90-4e94-b703-5fa5aa96be7b"}]}, {"file_name": "63ddf182-d8df-4a46-9231-6090272d2afa.mirbase21.mirnas.quantification.txt", "data_format": "TXT", "access": "open", "file_id": "eb73f062-1fe7-484e-ab22-856d2fd936ad", "data_category": "Transcriptome Profiling", "file_size": 50154, "cases": [{"project": {"project_id": "TCGA-LUAD"}, "case_id": "beb0025-74e2-451b-93b3-86f82df43573"}]}
```

Click on the tab , and check all the following items, then click on the JSON tab to download the case ids for the files.

# Part1: Data download, integration and preprocess

```
XINs-MacBook-Pro:miRNA xin$ python -3 ../../src/parse_file_case_id.py
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:16: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import multiarray
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:34: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import umath
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/_internal.py:204: DeprecationWarning: Overriding __e
q__ blocks inheritance of __hash__ in 3.x
  class dummy_ctype(object):
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/lib/mixins.py:63: DeprecationWarning: Overriding __eq__ b
```

Parse file case id by parse\_file\_case\_id.py



	A	B	C	D	E
1	file_id	case_id			
2	52792628-b45a2f8140-8f90-4e94-b703-5fa5aa96be7b				
3	eb73f062-1fcbeb0025-74e2-451b-93b3-86f82df43573				
4	a51d1524-3c82476d2d-e403-4f6b-8dd6-cc84e3329478				
5	143f51d7-4763bd2175-4b7c-44c9-aef3-9efc8f79837b				
6	824744e2-be0fdb152-25d2-404b-bec7-a43ece381f5b				
7	ee17ab5e-df27fceec1-3298-4cdd-a4e6-8f5cf34604f0				
8	951e488e-74f17d3463-af6c-42c2-a8e9-fae45f69286f				
9	2e83bcf1-54ebcba7f2-ce13-4bae-97cd-91a6b1dc465				
10	1ad0524e-4beb33753-9d38-4368-9033-3e55f129d00d				
11	e61f2ed3-54294cb595-0907-44c7-bbef-985a27c1e6e2				
12	56f851eb-ec9fcddcae-676e-4071-93c3-23d2d3ab0c00				
13	da50abb8-7l67dbe286-edb3-4d07-8f74-9e5254f01945				
14	6f3f3424-7d0d0d83d9-d558-4d38-977b-6f1b2471beda				
15	cef7be82-3babd6647b-d2eb-404d-8283-443e92addf55				
16	57f1ca14-6fce456a1f-6e2c-4b34-ad24-ccd614bd57c9				
17	fad7f43e-a3f30208a4c-ecb0-4b3c-a349-87e1c1f43e85				
18	8ff66bdc-8fd0f908c5-aafdf4b76-a2fb-39b7cc19bcf7				

# Part1: Data download, integration and preprocess

```
[XINs-MacBook-Pro:mirNA xin$ python -3 ../../src/request_meta.py
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/bindings/openssl/binding.py:14: DeprecationWarning: CObject type is not supported in 3.x. Please use capsule objects instead.
    from cryptography.hazmat.bindings.openssl import ffi, lib
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:150: DeprecationWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
    class DSAPrivateNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:189: DeprecationWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
    class DSAPublicNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:227: DeprecationWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
    class DSAPrivateNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/OpenSSL/crypto.py:463: DeprecationWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
    class X509Name(object):
```

sample_type	files_meta.tsv
cases.0.samples.0.portions.0.analytes.0.aliquots.0.aliquot_id	data_type
0.samples.0.tumor_descriptor	file_id
cases.0.submitter_id	cases.0.sample_id
cases.0.analytes.0.aliquots.0.aliquot_id	cases.0.case_id
cases.0.analytes.0.aliquots.0.aliquot_id	file_name
cases.0.analytes.0.aliquots.0.aliquot_id	cases.0.submitter_id
cases.0.analytes.0.aliquots.0.aliquot_id	cases.0.tissue_type
4aa894-0a42-4b20-9290-759a34e6f248	miRNA Expression Quantification Primary Tumor
0e22d215-92c2-435c-ad89-ee344ef88635	Transcriptome Profiling TCGA-68-A591
929556fd-6985-4ed9-96a5-e04a3ed3acbb	mirbase21.mirnas.quantification.txt
8856291a-fef5-4008-b6cc-636aa8795905	0fa76a1-4c3e-42e1-9d13-95aa3ae8e676
0e22d215-92c2-435c-ad89-ee344ef88635	TCGA-68-A591-01A
f4a61521-e301-4b9a-8958-e483136011d1	miRNA Expression Quantification Primary Tumor
50efcf8f-1e01-4f71-b9e4-f8491eb3c736	Transcriptome Profiling TCGA-22-5479
95532a5a-df6a-438a-a26d-7c4a1c241c9f	mirbase21.mirnas.quantification.txt
63157691-5ad8-4440-be22-c01a2614c9d0	fadd20a1-90ca-4d99-a128-9730d02a7bd6
50efcf8f-1e01-4f71-b9e4-50efcf8f-1e01-4f71-b9e4-	TCGA-22-5479-01A
f8491eb3c736	TCGA-22-5479-01A-31H-1948-13
8743ea2-a4e4-467d-8c31-49ec735bd4f2	miRNA Expression Quantification Primary Tumor
bb2acdaa-3a26-421a-a9e9-4ea3336a467b	Transcriptome Profiling TCGA-43-6771
8051f875-9875-4271-a757-4891db3c591a	mirbase21.mirnas.quantification.txt
12b448e3-2d58-92e1-dd484742cd99	b0818bee-c32c-405c-ba09-956091fbe09b
bb2acdaa-3a26-421a-50efcf8f-1e01-4f71-b9e4-	TCGA-43-6771-01A
a9e9-4ea3336a467b	TCGA-43-6771-01A-11H-1819-13
ad9be65-3dce-4b52-909c-99b07343d814	miRNA Expression Quantification Primary Tumor
be8861c3-48b9-4cb7-a9a5-d4e18449d3ce	Transcriptome Profiling TCGA-37-4141
7ae1a03b-cca0-a4d6-92b8-096e247c1b50	mirbase21.mirnas.quantification.txt
bcb18610-9e9d-4fbf-9698-82ceca1e935f	5d3d3918-7dab-4147-90a3-dc015beldfbe
be8861c3-48b9-4cb7-a9a5-5d3d3918-7dab-4147-90a3-dc015beldfbe	TCGA-37-4141-01A
d4e18449d3ce	TCGA-37-4141-01A-02T-1557-13
4eededad-03b6-4072-b518-b6760da4e656	miRNA Expression Quantification Primary Tumor
0435e383-6ec2-46c4-ba15-713a2447b0d1	Transcriptome Profiling TCGA-NC-A5HH
37be9876-e459-48e9-b9ba-b47d02c17ae7	mirbase21.mirnas.quantification.txt
3ac19224-4535-4f04-bb3d-7de9025542c2	b5865fle-604d-487d-8a19-dc0c887203ed
0435e383-6ec2-46c4-3ac19224-4535-4f04-bb3d-7de9025542c2	TCGA-NC-A5HH-01A
ba15-713a2447b0d1	TCGA-NC-A5HH-01A-11H-A26V-13
e15c3854-997c-4197-877f-99ac7695cb3a	miRNA Expression Quantification Primary Tumor
f7e9016d-2d0b-40d1-a330-09aae95684813	Transcriptome Profiling TCGA-22-5478
69c2dd2f-90f4-4f79-9d18-577ad2d53dfd	mirbase21.mirnas.quantification.txt
9fc48d9d-cfb3-4c04-b918-878c05c57a35	4daf4a91-bc36-40c8-8fcfa-ea61b6706775
f7e9016d-2d0b-40d1-69c2dd2f-90f4-4f79-9d18-577ad2d53dfd	TCGA-22-5478-01A
a330-09aae95684813	TCGA-22-5478-01A-01T-1634-13
e824558d-7170-4a3a-b483-e9fb02d9cf16	miRNA Expression Quantification Primary Tumor
07fa22d0-99b1-4364-9b93-d937005a2416	Transcriptome Profiling TCGA-22-1011
7df58443-5b11-4987-8227-6d809e8305af	mirbase21.mirnas.quantification.txt
ad8d4ef5-5c51-4342-983d-ef912fda745a	18058bf8-a08-410f-8e94-7c2b45a856b7
07fa22d0-99b1-4364-9b93-18058bf8-a08-410f-8e94-7c2b45a856b7	TCGA-22-1011-01A
d937005a2416	TCGA-22-1011-01A-01T-1557-13

primary_site	cases_meta.tsv
submitter_aliquot_ids.2	submitter_submitter_aliquot_ids.28
submitter_aliquot_ids.25	submitter_submitter_aliquot_ids.29
submitter_aliquot_ids.26	submitter_submitter_aliquot_ids.30
submitter_aliquot_ids.27	submitter_submitter_aliquot_ids.31
submitter_aliquot_ids.28	submitter_submitter_aliquot_ids.32
submitter_aliquot_ids.29	submitter_submitter_aliquot_ids.33
submitter_aliquot_ids.30	submitter_submitter_aliquot_ids.34
submitter_aliquot_ids.31	submitter_submitter_aliquot_ids.35
submitter_aliquot_ids.32	submitter_submitter_aliquot_ids.36
submitter_aliquot_ids.33	submitter_submitter_aliquot_ids.37
submitter_aliquot_ids.34	submitter_submitter_aliquot_ids.38
submitter_aliquot_ids.35	submitter_submitter_aliquot_ids.39
submitter_aliquot_ids.36	submitter_submitter_aliquot_ids.40
submitter_aliquot_ids.37	submitter_submitter_aliquot_ids.41
submitter_aliquot_ids.38	submitter_submitter_aliquot_ids.42
submitter_aliquot_ids.39	submitter_submitter_aliquot_ids.43
submitter_aliquot_ids.40	submitter_submitter_aliquot_ids.44
submitter_aliquot_ids.41	submitter_submitter_aliquot_ids.45
submitter_aliquot_ids.42	submitter_submitter_aliquot_ids.46
submitter_aliquot_ids.43	submitter_submitter_aliquot_ids.47
submitter_aliquot_ids.44	submitter_submitter_aliquot_ids.48
submitter_aliquot_ids.45	submitter_submitter_aliquot_ids.49
submitter_aliquot_ids.46	submitter_submitter_aliquot_ids.50
portion_ids.6	portion_ids.7
portion_ids.7	portion_ids.8
portion_ids.8	portion_ids.9
portion_ids.9	portion_ids.10
portion_ids.10	portion_ids.11
portion_ids.11	portion_ids.12
portion_ids.12	portion_ids.13
portion_ids.13	portion_ids.14
portion_ids.14	portion_ids.15
portion_ids.15	portion_ids.16
portion_ids.16	portion_ids.17
portion_ids.17	portion_ids.18
portion_ids.18	portion_ids.19
portion_ids.19	portion_ids.20
portion_ids.20	portion_ids.21
portion_ids.21	portion_ids.22
portion_ids.22	portion_ids.23
portion_ids.23	portion_ids.24
portion_ids.24	portion_ids.25
portion_ids.25	portion_ids.26
portion_ids.26	portion_ids.27
portion_ids.27	portion_ids.28
portion_ids.28	portion_ids.29
portion_ids.29	portion_ids.30
portion_ids.30	portion_ids.31
portion_ids.31	portion_ids.32
portion_ids.32	portion_ids.33
portion_ids.33	portion_ids.34
portion_ids.34	portion_ids.35
portion_ids.35	portion_ids.36
portion_ids.36	portion_ids.37
portion_ids.37	portion_ids.38
portion_ids.38	portion_ids.39
portion_ids.39	portion_ids.40
portion_ids.40	portion_ids.41
portion_ids.41	portion_ids.42
portion_ids.42	portion_ids.43
portion_ids.43	portion_ids.44
portion_ids.44	portion_ids.45
portion_ids.45	portion_ids.46
portion_ids.46	portion_ids.47
portion_ids.47	portion_ids.48
portion_ids.48	portion_ids.49
portion_ids.49	portion_ids.50
portion_ids.50	portion_ids.51
portion_ids.51	portion_ids.52
portion_ids.52	portion_ids.53
portion_ids.53	portion_ids.54
portion_ids.54	portion_ids.55
portion_ids.55	portion_ids.56
portion_ids.56	portion_ids.57
portion_ids.57	portion_ids.58
portion_ids.58	portion_ids.59
portion_ids.59	portion_ids.60
portion_ids.60	portion_ids.61
portion_ids.61	portion_ids.62
portion_ids.62	portion_ids.63
portion_ids.63	portion_ids.64
portion_ids.64	portion_ids.65
portion_ids.65	portion_ids.66
portion_ids.66	portion_ids.67
portion_ids.67	portion_ids.68
portion_ids.68	portion_ids.69
portion_ids.69	portion_ids.70
portion_ids.70	portion_ids.71
portion_ids.71	portion_ids.72
portion_ids.72	portion_ids.73
portion_ids.73	portion_ids.74
portion_ids.74	portion_ids.75
portion_ids.75	portion_ids.76
portion_ids.76	portion_ids.77
portion_ids.77	portion_ids.78
portion_ids.78	portion_ids.79
portion_ids.79	portion_ids.80
portion_ids.80	portion_ids.81
portion_ids.81	portion_ids.82
portion_ids.82	portion_ids.83
portion_ids.83	portion_ids.84
portion_ids.84	portion_ids.85
portion_ids.85	portion_ids.86
portion_ids.86	portion_ids.87
portion_ids.87	portion_ids.88
portion_ids.88	portion_ids.89
portion_ids.89	portion_ids.90
portion_ids.90	portion_ids.91
portion_ids.91	portion_ids.92
portion_ids.92	portion_ids.93
portion_ids.93	portion_ids.94
portion_ids.94	portion_ids.95
portion_ids.95	portion_ids.96
portion_ids.96	portion_ids.97
portion_ids.97	portion_ids.98
portion_ids.98	portion_ids.99
portion_ids.99	portion_ids.100
portion_ids.100	portion_ids.101
portion_ids.101	portion_ids.102
portion_ids.102	portion_ids.103
portion_ids.103	portion_ids.104
portion_ids.104	portion_ids.105
portion_ids.105	portion_ids.106
portion_ids.106	portion_ids.107
portion_ids.107	portion_ids.108
portion_ids.108	portion_ids.109
portion_ids.109	portion_ids.110
portion_ids.110	portion_ids.111
portion_ids.111	portion_ids.112
portion_ids.112	portion_ids.113
portion_ids.113	portion_ids.114
portion_ids.114	portion_ids.115
portion_ids.115	portion_ids.116
portion_ids.116	portion_ids.117
portion_ids.117	portion_ids.118
portion_ids.118	portion_ids.119
portion_ids.119	portion_ids.120
portion_ids.120	portion_ids.121
portion_ids.121	portion_ids.122
portion_ids.122	portion_ids.123
portion_ids.123	portion_ids.124
portion_ids.124	portion_ids.125
portion_ids.125	portion_ids.126
portion_ids.126	portion_ids.127
portion_ids.127	portion_ids.128
portion_ids.128	portion_ids.129
portion_ids.129	portion_ids.130
portion_ids.130	portion_ids.131
portion_ids.131	portion_ids.132
portion_ids.132	portion_ids.133
portion_ids.133	portion_ids.134
portion_ids.134	portion_ids.135
portion_ids.135	portion_ids.136
portion_ids.136	portion_ids.137
portion_ids.137	portion_ids.138
portion_ids.138	portion_ids.139
portion_ids.139	portion_ids.140
portion_ids.140	portion_ids.141
portion_ids.141	portion_ids.142
portion_ids.142	portion_ids.143
portion_ids.143	portion_ids.144
portion_ids.144	portion_ids.145
portion_ids.145	portion_ids.146
portion_ids.146	portion_ids.147
portion_ids.147	portion_ids.148
portion_ids.148	portion_ids.149
portion_ids.149	portion_ids.150
portion_ids.150	portion_ids.151
portion_ids.151	portion_ids.152
portion_ids.152	portion_ids.153
portion_ids.153	portion_ids.154
portion_ids.154	portion_ids.155
portion_ids.155	portion_ids.156
portion_ids.156	portion_ids.157
portion_ids.157	portion_ids.158
portion_ids.158	portion_ids.159
portion_ids.159	portion_ids.160
portion_ids.160	portion_ids.161
portion_ids.161	portion_ids.162
portion_ids.162	portion_ids.163
portion_ids.163	portion_ids.164
portion_ids.164	portion_ids.165
portion_ids.165	portion_ids.166
portion_ids.166	portion_ids.167
portion_ids.167	portion_ids.168
portion_ids.168	portion_ids.169
portion_ids.169	portion_ids.170
portion_ids.170	portion_ids.171
portion_ids.171	portion_ids.172
portion_ids.172	portion_ids.173
portion_ids.173	portion_ids.174
portion_ids.174	portion_ids.175
portion_ids.175	portion_ids.176
portion_ids.176	portion_ids.177
portion_ids.177	portion_ids.178
portion_ids.178	portion_ids.179
portion_ids.179	portion_ids.180
portion_ids.180	portion_ids.181
portion_ids.181	portion_ids.182
portion_ids.182	portion_ids.183
portion_ids.183	portion_ids.184
portion_ids.184	portion_ids.185
portion_ids.185	portion_ids.186
portion_ids.186	portion_ids.187
portion_ids.187	portion_ids.188
portion_ids.188	portion_ids.189
portion_ids.189	portion_ids.190
portion_ids.190	portion_ids.191
portion_ids.191	portion_ids.192
portion_ids.192	portion_ids.193
portion_ids.193	portion_ids.194
portion_ids.194	portion_ids.195
portion_ids.195	portion_ids.196
portion_ids.196	portion_ids.197
portion_ids.197	portion_ids.198
portion_ids.198	portion_ids.199
portion_ids.199	portion_ids.200
portion_ids.200	portion_ids.201
portion_ids.201	portion_ids.202
portion_ids.202	portion_ids.203
portion_ids.203	portion_ids.204
portion_ids.204	portion_ids.205
portion_ids.205	portion_ids.206
portion_ids.206	portion_ids.207
portion_ids.207	portion_ids.208
portion_ids.208	portion_ids.209
portion_ids.209	portion_ids.210
portion_ids.210	portion_ids.211
portion_ids.211	portion_ids.212
portion_ids.212	portion_ids.213
portion_ids.213	portion_ids.214
portion_ids.214	portion_ids.215
portion_ids.215	portion_ids.216
portion_ids.216	portion_ids.217
portion_ids.217	portion_ids.218
portion_ids.218	portion_ids.219
portion_ids.219	portion_ids.220
portion_ids.220	portion_ids.221
portion_ids.221	portion_ids.222
portion_ids.222	portion_ids.223
portion_ids.223	portion_ids.224
portion_ids.224	portion_ids.225
portion_ids.225	portion_ids.226
portion_ids.226	portion_ids.227
portion_ids.227	portion_ids.228
portion_ids.228	portion_ids.229
portion_ids.229	portion_ids.230
portion_ids.230	portion_ids.231
portion_ids.231	portion_ids.232
portion_ids.232	portion_ids.233
portion_ids.233	portion_ids.234
portion_ids.234	portion_ids.235
portion_ids.235	portion_ids.236
portion_ids.236	portion_ids.237
portion_ids.237	portion_ids.238
portion_ids.238	portion_ids.239
portion_ids.239	portion_ids.240
portion_ids.240	portion_ids.241
portion_ids.241	portion_ids.242
portion_ids.242	portion_ids.243
portion_ids.243	portion_ids.244
portion_ids.244	portion_ids.245
portion_ids.245	portion_ids.246
portion_ids.246	portion_ids.247
portion_ids.247	portion_ids.248
portion_ids.248	portion_ids.249
portion_ids.249	portion_ids.250
portion_ids.250	portion_ids.251
portion_ids.251	portion_ids.252
portion_ids.252	portion_ids.253
portion_ids.253	portion_ids.254
portion_ids.254	portion_ids.255
portion_ids.255	portion_ids.256
portion_ids.256	portion_ids.257
portion_ids.257	portion_ids.258
portion_ids.258	portion_ids.259
portion_ids.259	portion_ids.260
portion_ids.260	portion_ids.261
portion_ids.261	portion_ids.262
portion_ids.262	portion_ids.263
portion_ids.263	portion_ids.264
portion_ids.264	portion_ids.265
portion_ids.265	portion_ids.266
portion_ids.266	portion_ids.267
portion_ids.267	portion_ids.268

# Part1: Data download, integration and preprocess

```
n_labels = df['label'].value_counts()  
print(n_labels)  
print(df.shape)
```

```
2.0      1104  
4.0      1081  
22.0     1000  
0.0      691  
13.0     550  
6.0      532  
23.0     515  
29.0     500  
5.0      499  
1.0      459  
20.0     452  
12.0     450  
17.0     418  
33.0     408  
16.0     309  
14.0     301  
7.0      231  
3.0      185  
21.0     179  
35.0     157  
11.0     131  
34.0     129  
28.0     119  
18.0     118  
19.0     117  
27.0     103
```

```
labels = {0: 0, 'Colon': 1, 'Breast': 2, 'Esophagus': 3, nan: 60, 'Kidney': 4, 'Prostate gland': 5,  
'Brain': 6, 'Adrenal gland': 7, 'Thymus': 8, 'Lymph nodes': 9,  
'Base of tongue': 10, 'Other and unspecified parts of tongue': 11, 'Skin': 12,  
'Corpus uteri': 13, 'Blood': 14, '0': 15, 'Cervix uteri': 16, 'Bladder': 17,  
'Retroperitoneum and peritoneum': 18, 'Larynx': 19, 'Stomach': 20, 'Pancreas': 21,  
'Bronchus and lung': 22, 'Thyroid gland': 23, 'Tonsil': 24, 'Rectum': 25,  
'Hypopharynx': 26, 'Hematopoietic and reticuloendothelial systems': 27,  
'Heart, mediastinum, and pleura': 28, 'Ovary': 29, 'Uterus, NOS': 30,  
'Other and ill-defined sites in lip, oral cavity and pharynx': 31,  
'Eye and adnexa': 32, 'Liver and intrahepatic bile ducts': 33,  
'Connective, subcutaneous and other soft tissues': 34, 'Testis': 35,  
'Rectosigmoid junction': 36, 'Other and unspecified parts of mouth': 37,  
'Other and ill-defined sites': 38, 'Floor of mouth': 39,  
'Other endocrine glands and related structures': 40, 'Lip': 41,  
'Small intestine': 42, 'Gum': 43, 'Oropharynx': 44,  
'Peripheral nerves and autonomic nervous system': 45,  
'Bones, joints and articular cartilage of other and unspecified sites': 46,  
'Unknown primary site': 47, 'Palate': 48,  
'Bones, joints and articular cartilage of limbs': 49,  
'Other and unspecified parts of biliary tract': 50, 'Meninges': 51,  
'Gallbladder': 52,  
'Spinal cord, cranial nerves, and other parts of central nervous system': 53,  
'Other and unspecified male genital organs': 54}
```

- Generate the miRNA matrix for all the files with labeled normal or tumor.  
The miRNA seq that comes from totally so many different cancers based on  
“primary site” in cases meta
- Remove some primary site type which counts are less than 100, and then  
only 26 primary site left
- Normal tissue is labeled with 0, and other primary sites labeled the different  
values based on dictionary in gen\_miRNA\_matrix.py
- Finally, there still have over 10,000 samples

# Part2: Prediction using KNN

hsa-mir-941	hsa-mir-941	hsa-mir-941	hsa-mir-942	hsa-mir-943	hsa-mir-944	hsa-mir-95	hsa-mir-950	hsa-mir-96	hsa-mir-98	hsa-mir-99a	hsa-mir-99b	label	▼
0	0	0	13	0	0	4	0	1	84	762	75212	1	
0	0	0	13	0	2	94	0	79	530	12392	233865	2	
0	0	0	101	0	2078	44	0	92	338	782	88949	4	
0	0	0	72	0	3	18	0	83	374	9898	153229	5	
0	0	0	24	0	0	5	0	2	24	461	871177	6	
0	0	0	3	0	1	1	0	27	114	395	37435	7	
0	0	0	35	1	600	5	0	80	97	490	12794	6	
0	0	0	18	0	0	11	0	240	211	885	72173	12	
0	0	0	43	2	0	36	0	157	185	2228	146318	13	
0	0	0	60	0	975	88	0	182	174	4536	43652	14	
0	0	0	10	0	38	8	0	11	92	3055	32007	0	
0	0	0	4	0	0	1	0	94	133	3079	52183	16	
0	0	0	91	1	310	3	0	150	133	442	29625	7	
0	0	0	15	0	2	6	0	21	114	280	42735	6	
0	0	0	42	0	837	30	0	56	253	726	48645	4	
0	0	0	8	0	0	2	0	5	256	3483	261715	17	

```
# split the data to train and test set
X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.3, random_state=0)

#standardize the data.
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

- Read data from miRNA matrix
- Scatter data to train set and test set 7:3

# Part2: Prediction using KNN

```
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
num_features=800
pca = PCA(n_components=num_features, whiten=True)
pca.fit(X_train)
X_train = pd.DataFrame(pca.transform(X_train))
X_test = pd.DataFrame(pca.transform(X_test))
print("PCA reduction")
print("feature numbers: "+str(num_features))
#print("explained_variance_ratio_: \n"+str(pca.explained_variance_ratio_))
print("explained_variance_ratio_.cumsum: \n"+str(pca.explained_variance_ratio_.cumsum()))
```

```
PCA reduction
feature numbers: 800
explained_variance_ratio_.cumsum:
[0.08024552 0.11213829 0.14088122 0.16403317 0.18027623 0.19439016
 0.20674242 0.21807625 0.22671984 0.234468 0.2417881 0.24870134
 0.25542148 0.26137246 0.26716042 0.27258638 0.27793182 0.28295968
 0.28767314 0.2920368 0.29624207 0.30035979 0.30441073 0.30826549
 0.31186178 0.31539722 0.3188568 0.32218002 0.32537988 0.32846721
 0.3314421 0.3342966 0.33710044 0.33986321 0.34253281 0.34517575
 0.34776262 0.35033472 0.35283313 0.35530327 0.35770834 0.36010286
 0.3624359 0.36468992 0.366889 0.36907654 0.37121741 0.37329379
 0.37536765 0.37741505 0.37944168 0.38144251 0.38343012 0.38540043
- - - - -
```

- Feature selection using PCA and t-SNE
- Visualize the result – we can see that reducing dimension from 1883 to 800 has already been able to keep 89% information
- We can not use the higher dimension because our sample number is 8788 after data processing

```
.....
0.87386774 0.87418761 0.87450662 0.87482541 0.87514312 0.87546015
0.87577579 0.87609111 0.87640541 0.87671878 0.87703196 0.87734436
0.87765576 0.87796593 0.87827549 0.87858431 0.87889263 0.8792
0.87950651 0.87981199 0.88011685 0.88042108 0.88072333 0.88102454
0.88132522 0.88162546 0.88192437 0.8822216 0.88251842 0.88281364
0.88310788 0.88340136 0.8836945 0.88398741 0.88427956 0.88457093
0.88486063 0.88515005 0.88543808 0.88572572 0.88601299 0.88629756
0.88658203 0.8868655 0.88714781 0.88742977 0.88771032 0.88798906
0.88826589 0.88854103 0.88881472 0.88908791 0.88936033 0.88963194
0.88990166 0.89017089]
```

# Part2: Prediction using KNN

```
print("t-SNE evaluation")
tsne = TSNE(n_components=2, verbose=1, perplexity=40, n_iter=250)
new = tsne.fit_transform(X_train)
features = pd.DataFrame(new)

t-SNE evaluation
[t-SNE] Computing 121 nearest neighbors...
[t-SNE] Indexed 6151 samples in 0.006s...
[t-SNE] Computed neighbors for 6151 samples in 0.869s...
[t-SNE] Computed conditional probabilities for sample 1000 / 6151
[t-SNE] Computed conditional probabilities for sample 2000 / 6151
[t-SNE] Computed conditional probabilities for sample 3000 / 6151
[t-SNE] Computed conditional probabilities for sample 4000 / 6151
[t-SNE] Computed conditional probabilities for sample 5000 / 6151
[t-SNE] Computed conditional probabilities for sample 6000 / 6151
[t-SNE] Computed conditional probabilities for sample 6151 / 6151
[t-SNE] Mean sigma: 0.090937
```

```
print(features)
```

	0	1
0	-0.972481	-0.445210
1	0.697620	-2.116041
2	-0.782807	-0.510700
3	0.611465	2.169794
4	-0.048844	2.251833
5	-0.007524	-1.655136
6	-0.291251	-0.689894
7	0.017664	-1.646448
8	-0.730924	-0.151445
9	-0.947784	1.730283
10	-1.055320	-2.054043
11	-1.086948	0.247150
12	0.656100	2.440754
13	-0.108890	-3.028151
14	-0.472456	-0.835887
15	-0.244564	-0.809207
16	0.128819	1.124375
17	0.328129	1.508422
18	-0.805383	-2.538891
19	-0.571301	-4.143396
20	-1.211597	0.818011
21	0.162854	-1.594920
22	-0.197186	-2.877308
23	1.445303	-0.139750
24	0.598879	2.575904
25	-0.526559	0.121936
26	-0.633962	0.442991
27	1.095082	-1.749704
28	0.636407	1.886242
29	-0.506113	-4.005802
...	...	...

- Feature selection using PCA and t-SNE
- Visualize the result – we can see that reducing dimension from 1883 to 800 has already been able to keep 89% information

# Part2: Prediction using KNN

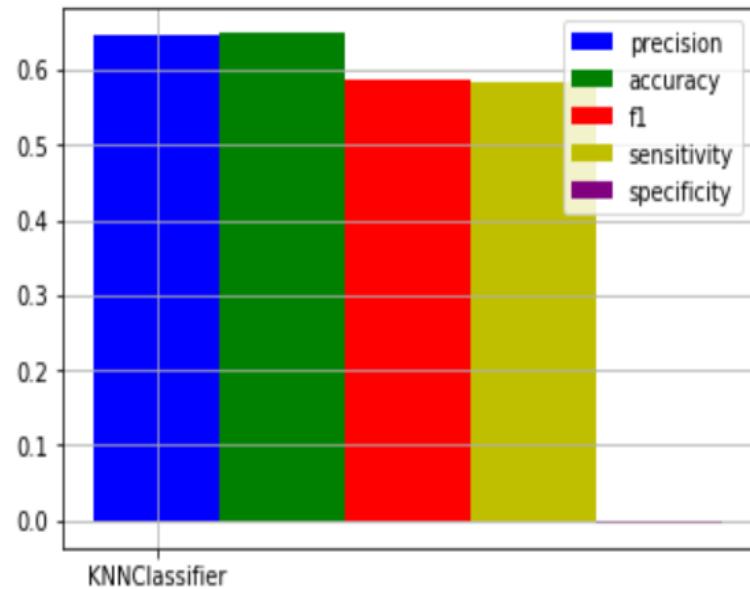
```
models = {
    'KNeighborsClassifier': KNeighborsClassifier(),
}
tuned_parameters = {
    'KNeighborsClassifier': {'n_neighbors': [3, 5, 7, 10]},
}

clf = GridSearchCV(models[key], tuned_parameters[key], scoring=None, refit=True, cv=10)
clf.fit(X_train,y_train)
y_test_predict = clf.predict(X_test)
precision = precision_score(y_test, y_test_predict, average='micro')
accuracy = accuracy_score(y_test, y_test_predict)
f1 = f1_score(y_test, y_test_predict, average='micro')
recall = recall_score(y_test, y_test_predict, average='micro')
specificity = specificity_score(y_test, y_test_predict)
scores[key] = [precision,accuracy,f1,recall,specificity]
# return the best model which K hyper-parameter can get the highest performance
best_model = clf.best_estimator_
```

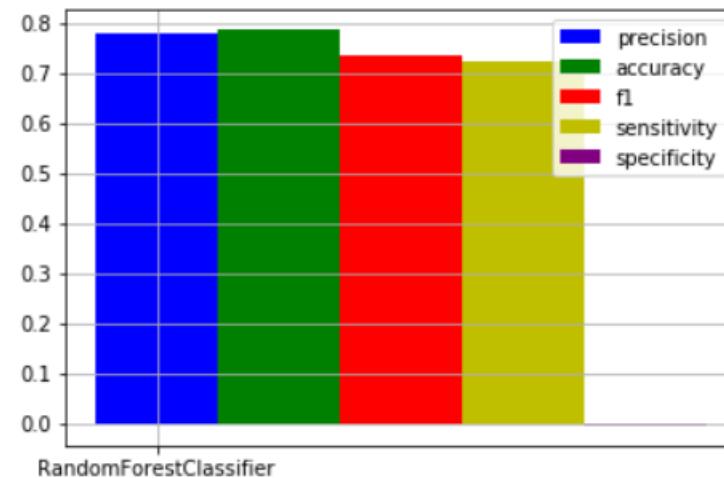
- Hyper-parameter tuning – we select  $K = \{3, 5, 7, 10\}$  to try to get the best  $K$  using grid search CV
- Score average set ‘micro’ for multiclass, which will return the total ratio

# Part2: Prediction using KNN

```
['KNNClassifier']
```



```
['RandomForestClassifier']
```



```
In [6]: print(scores)
```

```
{'KNNClassifier': [0.6448927653055788, 0.6485780615206036, 0.5867057788654338, 0.584273255563632, -0.0036629130058161118]}
```

- Predict and output evaluation metrics and ROC curve

<https://github.com/appleshine77/GDC-data-lab.git>