EE 542 – Laboratory Assignment
Instructor: Young H. Cho
T.A.: Yue Shi
Due date: October 10 at 11:59pm

The git repo for all the codes provided in this lab: https://github.com/yuesOctober/GDCproject/tree/yue

Download the repo:
git clone https://github.com/yuesOctober/GDCproject.git

Basic git command : https://confluence.atlassian.com/bitbucketserver/basic-git-commands-776639767.html

GDC Data Lab:

In this lab, you will learn

1. **How to download, integrate, and preprocess files related to a particular disease type, and how to use the data obtained.**
2. **As an example, you will go through the entire process to get the miRNA files, and the related file metadata, case metadata to the disease Liver Hepatocellular Carcinoma**
3. **You will apply the machine learning package to the miRNA matrix extracted to detect normal/ cancer samples.**

What to turn in:

Go through the entire tutorial and do the Part 1 and Part 2 with the disease type: Lung Squamous Cell Carcinoma. In Part2, try a different model other than the one provided in the sample code and plot the ROC curve for the models.

Extra Credit: Explore the Gene Expression Quantification Data.

Submission guideline:

Each team should create a github repo and provide the link to your repo for code and slide submission. You need have a readme file explaining how to run your source codes. For video demo submission, you need show the steps to run your code and explain. Only one submission per team is needed.

Part 1: Data download, integration and preprocess.

1. Introduction to GDC data:

Read the document below to get a sense of GDC data.

https://gdc.cancer.gov/about-data

Biomarker Data:

Data Category	Data Type
DNA Methylation	Methylation Beta Value
Simple Nucleotide Variation	Annotated Somatic Mutation
	Raw Simple Somatic Mutation
	Aggregated Somatic Mutation
	Masked Somatic Mutation
Transcriptome Profile	Gene Expression Quantification
	Isoform Expression Quanfitication
	miRNA expression Quantification

2. **Example: Downloading miRNA files of Disease: Liver Hepatocellular Carcinoma**

miRNA Expression Quantification is a table that associates miRNA IDs with read count and a normalized count in reads-per-million-miRNA-mapped. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/

Download **Expression Quantification data:** miRNA sequence data

1. Go to the data portal *https://portal.gdc.cancer.gov/repository*, on the left side there are two tabs : ***Files*** and ***Cases***
2. Click ***Cases*** and select a disease type: **Liver Hepatocellular Carcinoma**
3. Click **Files** and select
   Data Category: **Transcriptome Profiling**
   Data type : **miRNA Expression Quantification**
   Experimental Strategy: **miRNA-Seq**

You will see 373 cases and 425 files. That means there are duplicates for some cases. Also in those cases, there are some normal cases without cancer.

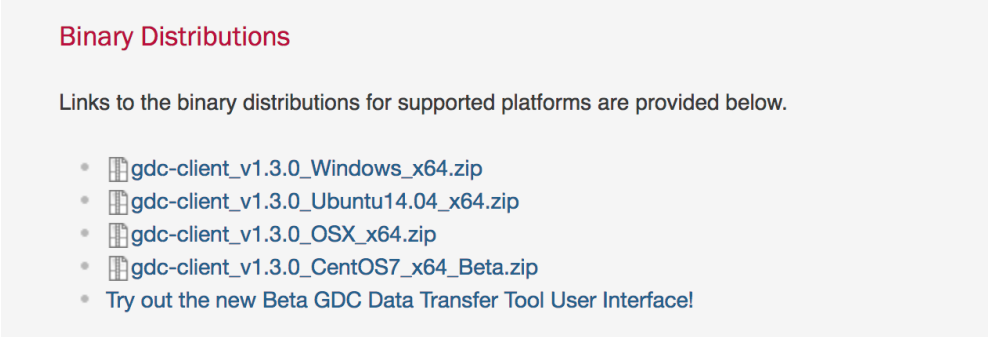
4. Click on the **Manifest download.** This will download the manifest file for use with GDC data transfer tool.

   The Manifest file contains the id, filename, md5, size and patient state.

id	filename	md5	size	state
baa65cc1-acb7-46c0-b68b-ce11600b476d	0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt	1a94e462dd73d4bbb8a2866e09d568e6	50268	live
593d4a08-a05e-42c8-9440-4176fbf177fe	455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt	66e27e0fbbeb4fd0de482c0e1b7194d8	50261	live
9a7a5f6e-5b59-48d4-8ad6-608ddd5e739e	8c8e40d1-f62f-4d9d-9435-ef0bc4fcd64e.mirbase21.mirnas.quantification.txt	86650cd63bbd18211f37d42751974877	50137	live
1e78c8a5-aed7-4cf7-8c99-ba2220053948	7bb700da-edf4-4831-80ef-8d252f590257.mirbase21.mirnas.quantification.txt	4ce6c76f6af6ff9684d4698346ea8895	50301	live
517095c1-a30d-4582-9f41-fcb9ddd251e1	f7f8c1ac-96cb-49bf-a485-dc8404105191.mirbase21.mirnas.quantification.txt	ab7e824cf406d4af257c95535c456b50	50046	live
9b23f8cb-6b59-4040-b7c8-ba4fa08eba55	466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt	aec98f0de51afae776f88218f9c6676d	50206	live
0b74f41b-1771-4f42-8181-ca6fc7686b5c	6bee6719-9ee9-4561-8c59-1667f2632d52.mirbase21.mirnas.quantification.txt	e3eca02afce43633544f433cbdb4dfd3	50358	live
963ea266-2577-425f-b47d-168d6c8d7c72	820f4603-0895-4741-865c-fdd98fbcb4fb.mirbase21.mirnas.quantification.txt	06f2c8199f2ff690f6191d14e89bc6ac	50133	live
9a1850a4-7050-4c64-8d3d-50d040cb7c89	bb405508-0f6a-4353-8c87-887be99855d5.mirbase21.mirnas.quantification.txt	fd40688c8c81efa0c88cdb2b68d051de	50402	live
74026969-eab8-44f1-8746-d89d5a450ba1	45da1c01-0316-4dbf-939b-4a758fd7e5e7.mirbase21.mirnas.quantification.txt	f20d3488568b2446f558fd50e0f9e7c4	50294	live
772f0a50-c019-4d85-a1c9-6534ff57f482	f7332a1d-ba16-44cd-b6c8-2639fdd568bf.mirbase21.mirnas.quantification.txt	a09f9536d861e0ae81ff46ec3597b2ed	50361	live

5. Data transfer tool Download:
   https://gdc.cancer.gov/access-data/gdc-data-transfer-tool
   Download the version according to your OS type.
   Command line to **download** and **unzip** a **OSX** version:
   Download:
   wget -c -t 0 https://gdc.cancer.gov/files/public/file/gdc-client_v1.3.0_OSX_x64.zip
   Unzip:
   Unzip gdc-client_v1.3.0_OSX_x64.zip

   Note: For other versions, just replace with the corresponding OS version file name.



System Recommendations

The system recommendations for using the GDC Data Transfer Tool are as follows:

- **OS: Linux (Ubuntu 14.x or later), OS X (10.9 Mavericks or later), or Windows (7 or later)**
- **CPU: At least eight 64-bit cores, Intel or AMD**
- **RAM: At least 8 GiB**
- **Storage: Enterprise-class storage system capable of at least 1 Gb/s (gigabit per second) write throughput and sufficient free space for BAM files.**

6. Download the files with gdc-client tool :
   a. make a directory for the data:
      mkdir live_miRNA
      cd live_miRNA
   b. Download with gdc-client.
      ./<path-to-gdc-client>/gdc-client download –m <path-to-manifest-file>
      e.g.
   ./~/Downloads/gdc-client –m ~/Downloads/gdc_manifest.2018-08-23.txt

   After successful downloads, you will see

```
100% [##############################################] Time: 0:00:00  69.62 kB/s
100% [##############################################] Time: 0:00:00  74.38 kB/s
100% [##############################################] Time: 0:00:00  66.29 kB/s
100% [##############################################] Time: 0:00:00  73.63 kB/s
100% [##############################################] Time: 0:00:00  67.27 kB/s
100% [##############################################] Time: 0:00:00  69.77 kB/s
Successfully downloaded: 425
```

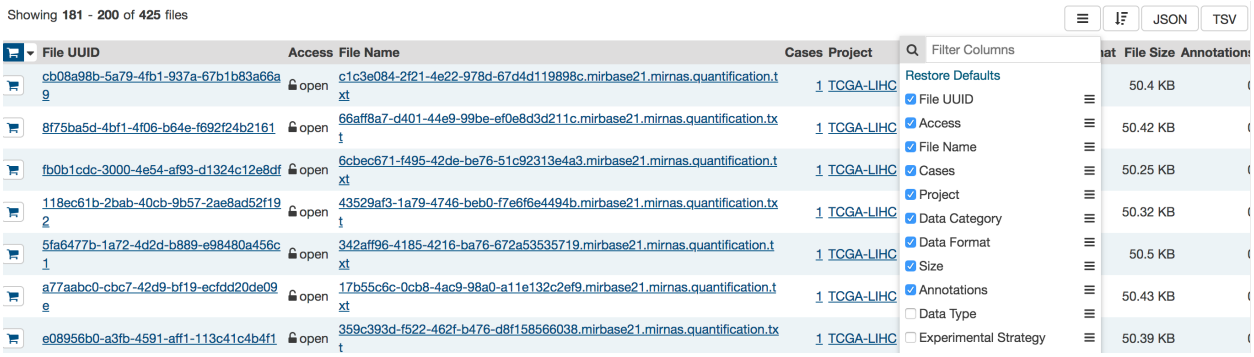
7. Check the successful download:
   Since large volumes of data are downloaded, it is important to check the file integrity. You could use the md5 checksum to check the integrity of downloaded files.
   Run the code: python3 check.py
   A sample python 3 code **check.py** is provided.

8. If some files fail download, use the following command:
   ./<path-to-gdc-client>/gdc-client download <id>
   e.g.
   ../gdc-client download fa63ce14-b9b5-4041-9df7-3b86ba9ede16

9. Once we get the biomarker files. We also need get the case ids related to the files .
   This is because we need correlate the biomarker files with the corresponding case clinical/ biospecimen files.
   Here we need to write some python codes to extract all the file_ids and the corresponding case_ids for future use.
   Get the cases related to the files:

The code ***parse_file_case_id.py*** is provided.

Click on the tab ☰, and check all the following items, then click on the **JSON** tab. It will download the case ids for the files.



Screenshot of a downloaded file:

```
[{
  "file_name": "0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt",
  "data_format": "TXT",
  "access": "open",
  "file_id": "baa65cc1-acb7-46c0-b68b-ce11600b476d",
  "data_category": "Transcriptome Profiling",
  "file_size": 50268,
  "cases": [
    {
      "project": {
        "project_id": "TCGA-LIHC"
      },
      "case_id": "7bdc5f86-4d7d-4f1f-bc23-ab51fa9fb947"
    }
  ]
},{
  "file_name": "455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt",
  "data_format": "TXT",
  "access": "open",
  "file_id": "593d4a08-a05e-42c8-9440-4176fbf177fe",
  "data_category": "Transcriptome Profiling",
  "file_size": 50261,
  "cases": [
    {
      "project": {
        "project_id": "TCGA-LIHC"
      },
      "case_id": "801b1d2c-eb6f-4eef-a00b-83da939d755a"
    }
  ]
```


11. Get the meta data for the files and corresponding cases:
The source code: ***request_meta.py***
The fields for the files and cases:
File fields:
https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/#file-fields
case fields:
https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/#case-fields

Once we get the meta data for the miRNA files, we can see that some samples come from a normal solid tissue and some others come from tumor.


```
cases.0.samples.0.portions.0.analytes.0.aliquots.0.aliquot_id   data_type       cases.0.samples.0.sample_type   file_name       cases.0.samples.0.submitter_id
cases.0.samples.0.tissue_type   cases.0.samples.0.tumor_descriptor      file_id data_category   cases.0.submitter_id    cases.0.samples.0.sample_id     cases.
0.case_id       id      cases.0.samples.0.portions.0.analytes.0.aliquots.0.submitter_id
76d02c30-f4b0-4d7b-85cf-f022c14cf0ae    miRNA Expression Quantification Primary Tumor   0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt
        TCGA-ZS-A9CE-01A                baa65cc1-acb7-46c0-b68b-ce11600b476d    Transcriptome Profiling TCGA-ZS-A9CE    12076740-
d690-48ac-9630-5105ac239111     7bdc5f86-4d7d-4f1f-bc23-ab51fa9fb947    baa65cc1-acb7-46c0-b68b-ce11600b476d    TCGA-ZS-A9CE-01A-11R-A37G-13
d5c5ae8e-9f61-4977-ba2e-df4d2486edc4    miRNA Expression Quantification Primary Tumor   455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt
        TCGA-2Y-A9H4-01A                593d4a08-a05e-42c8-9440-4176fbf177fe    Transcriptome Profiling TCGA-2Y-A9H4    c53e6fd9-76e1-44c3-af36-
d1d0dfc2802a    801b1d2c-eb6f-4eef-a00b-83da939d755a    593d4a08-a05e-42c8-9440-4176fbf177fe    TCGA-2Y-A9H4-01A-11R-A38M-13
d62a0ef1-2a67-47e7-b468-2a00c25ebab4    miRNA Expression Quantification Primary Tumor   8c8e40d1-f62f-4d9d-9435-ef0bc4fcd64e.mirbase21.mirnas.quantification.txt
        TCGA-BC-A69I-01A                9a7a5f6e-5b59-48d4-8ad6-608ddd5e739e    Transcriptome Profiling TCGA-BC-A69I    e36d5a75-0bc8-480d-8458-
e23cb453c279    70aecc90-17f0-4064-8ebb-be81f9b1d7fd    9a7a5f6e-5b59-48d4-8ad6-608ddd5e739e    TCGA-BC-A69I-01A-11R-A310-13
d7fa37f0-47d0-4f67-a36b-f4dbbd1a63cb    miRNA Expression Quantification Solid Tissue Normal     7bb700da-
edf4-4831-80ef-8d252f590257.mirbase21.mirnas.quantification.txt TCGA-BC-A10R-11A                1e78c8a5-aed7-4cf7-8c99-ba2220053948    Transcriptome
Profiling       TCGA-BC-A10R    1ea00561-3d1d-49d0-a3f7-d34eb8fec235    0bf5bbd4-d9e8-42a6-9ab5-f2c174dec12c    1e78c8a5-aed7-4cf7-8c99-ba2220053948    TCGA-BC-
A10R-11A-11R-A130-13
6fa420dd-c227-4153-8c3e-f85a55340bd1    miRNA Expression Quantification Primary Tumor   f7f8c1ac-96cb-49bf-a485-dc8404105191.mirbase21.mirnas.quantification.txt
        TCGA-CC-A7IL-01A                517095c1-a30d-4582-9f41-fcb9ddd251e1    Transcriptome Profiling TCGA-CC-A7IL
1445b31d-812d-4ddf-8965-0a970b9ea562    b54df65f-8a6b-4405-9918-90d53418804c    517095c1-a30d-4582-9f41-fcb9ddd251e1    TCGA-CC-A7IL-01A-11R-A343-13
6cb46502-3901-4568-ac87-ab03e7615b03    miRNA Expression Quantification Primary Tumor   466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt
        TCGA-DD-A118-01A                9b23f8cb-6b59-4040-b7c8-ba4fa08eba55    Transcriptome Profiling TCGA-DD-A118
b1da5474-3d42-433c-918e-22cb5c353a13    7d914d07-f250-4c7a-8fa2-df14e708936e    9b23f8cb-6b59-4040-b7c8-ba4fa08eba55    TCGA-DD-A118-01A-11R-A130-13
8ca1ceb5-1122-4bd8-bf4f-615658f0ba1b    miRNA Expression Quantification Primary Tumor   6beee6719-9ee9-4561-8c59-1667f2632d52.mirbase21.mirnas.quantification.txt
        TCGA-DD-AACW-01A                0b74f41b-1771-4f42-8181-ca6fc7686b5c    Transcriptome Profiling TCGA-DD-AACW    32c40946-b067-48a5-8a7c-
cd1923bd6d60    67a00f5f-c753-48f9-bc24-8287f50ec776    0b74f41b-1771-4f42-8181-ca6fc7686b5c    TCGA-DD-AACW-01A-11R-A41H-13
607da7cc-3cb5-426c-9c82-fe2bead2b392    miRNA Expression Quantification Primary Tumor   820f4603-0895-4741-865c-fdd98fbcb4fb.mirbase21.mirnas.quantification.txt
```

12. Now we could generate the miRNA matrix for all the files with labeled normal or tumor. The miRNA seq that comes from tumor is labeled with 1, and normal tissue is labeled with 0. The source code: ***gen_miRNA_matrix.py***

file_id	hsa-let-7a-1	hsa-let-7a-2	hsa-let-7a-3	hsa-let-7b	hsa-let-7c	hsa-let-7d	hsa-let-7e	hsa-let-7f-1	hsa-let-7f-2	hsa-let-7g	hsa-let-7i	hsa-mir-1-1	hsa-mir-1-2	hsa-mir-100	hsa-mir-101
0032767c-7b	51942	52033	52046	19640	5011	2156	3274	54448	57507	4827	805	15	12	9382	19746
02b0dfe5-79	31753	31748	31951	15313	15896	1623	2427	19689	21152	4111	530	3	3	66403	35841
038e6b5d-0c	24925	24774	24853	7917	7301	509	780	4612	4727	1337	396	5	7	76365	61711
03cd9e0d-b4	113993	113202	114613	124412	12011	9231	33424	70109	71104	7155	2832	59	74	33209	31593
050a903a-79	124631	124247	123967	114136	37963	2965	2743	87933	88721	6022	1571	5	10	142364	67009
052a2b32-c3	118185	117792	118548	92063	38018	3168	9848	84238	86009	10212	2693	48	63	78041	120628
05a7b6b3-e2	74823	74216	75025	85169	2462	4272	3168	54479	56148	4929	1320	16	14	1442	33793
0622e467-b1	35132	35137	35450	21659	682	6970	2544	31782	32416	1940	1303	22	22	1213	17894
0820d171-84	191442	191969	192254	159411	11607	7246	8103	156434	158877	7974	2431	11	23	42237	57149
0859bada-df	33808	32925	33133	17699	17871	1262	1698	15302	16840	2360	641	17	6	42855	36139
08b92aea-9a	17536	17259	17602	16693	11146	1176	987	10310	10750	2901	827	2	7	20288	19015
0963590c-70	33214	34013	33763	39531	25407	1824	2951	11526	11658	4216	1009	11	16	93039	43647
097683f7-5d	52762	52603	53172	69877	11689	3002	5682	17186	17490	4598	2003	24	34	54655	96168
09df6fb8-c2e	49757	49391	49404	80340	2217	2936	6262	22251	22314	2725	1555	4	10	14426	26803
0a4e1dee-7b	82413	82164	83139	27041	16057	1459	12192	17563	18127	3674	1203	7	7	8108	153174
0b66effa-907	123155	122989	122917	69864	53468	2438	5362	87106	89686	8784	1778	24	24	68520	194575
0b74f41b-17	59143	58749	59842	34779	18478	8572	1386	54348	54435	3691	899	3	8	44267	52282
0ca2657a-87	31043	30699	31040	38303	5818	6475	1414	18351	18767	2483	1140	7	8	5287	16720
0cb6c6e1-d1	20252	20294	20582	27044	13034	1275	1283	7237	7559	2171	786	5	10	20812	75085
0d2141c1-c2	12397	12188	12386	15229	5522	1017	1310	4279	4628	1153	571	9	12	12007	6663
0d4f7e7b-dc	39383	38839	39335	59544	12698	1113	2610	19223	19254	2682	921	32	26	30951	67266
0e49336b-5f	26687	26495	26832	12188	5372	3082	520	21281	21813	4777	859	1	0	5760	38744
0f5c1ab7-f4b	21277	21083	20969	13011	4034	1421	2318	12194	13021	2747	695	5	3	40766	31558
0fa60991-5e	103984	103978	104897	119307	18425	5540	14553	57411	58649	5797	3534	13	9	59530	38280

	BTF	BTG	BTH	BTI	BTJ	BTK	
1	-95	hsa-mir-9500	hsa-mir-96	hsa-mir-98	hsa-mir-99a	hsa-mir-99b	label
2	15	0	177	250	1040	22624	1
3	40	0	12	197	8052	27649	1
4	57	0	200	43	5979	19477	1
5	49	0	93	752	2946	238146	1
6	29	0	7	771	5192	24778	1
7	63	0	8	469	8798	131700	0
8	38	0	2	513	338	38054	1
9	21	0	0	831	157	47211	1
10	49	0	311	1343	1546	73135	1
11	25	0	5	161	3650	24824	1
12	22	0	74	168	2462	19346	1
13	68	0	126	125	14658	75655	1
14	39	0	22	245	6248	154333	1
15	148	0	202	345	873	124912	1
16	86	0	97	197	8360	283988	1
17	33	0	8	482	11402	63169	0
18	5	0	0	649	7404	11177	1
19	25	0	44	396	1369	29653	1
20	10	0	3	93	5415	32795	0
21	21	0	5	73	1460	36546	1
22	7	0	16	109	3595	32811	1
23	64	0	106	201	3297	7269	1
24	15	0	22	106	2109	35195	1
25	201	0	148	687	4355	138078	1
26	27	0	5	246	8823	45085	0
27	9	0	17	217	1138	35981	1
28	158	0	49	366	1482	99628	1
29	14	0	1	299	2013	34278	1
30	29	0	5	221	7683	49631	0
31	18	0	1	127	686	13523	1
32	6	0	62	56	226	13816	1
33	12	0	5	443	7528	62275	0
34	44	0	22	471	10104	113164	0
35	66	0	706	361	2106	7150	1
36	7	0	101	121	4318	13144	1

Part 2: Apply Machine Learning Package (sklearn) to the above data.

Sample code provided: ***predict.py***

The steps are as following:

1. Data standardization.
2. Train and test data split.
3. Feature selection.
4. Model hyper-parameters tuning with cross validation
5. Model prediction with the best hyper-parameters
6. Evaluation: Precision, Sensitivity, Accuracy, F1-score, Specificity

The result is shown in Figure 1.

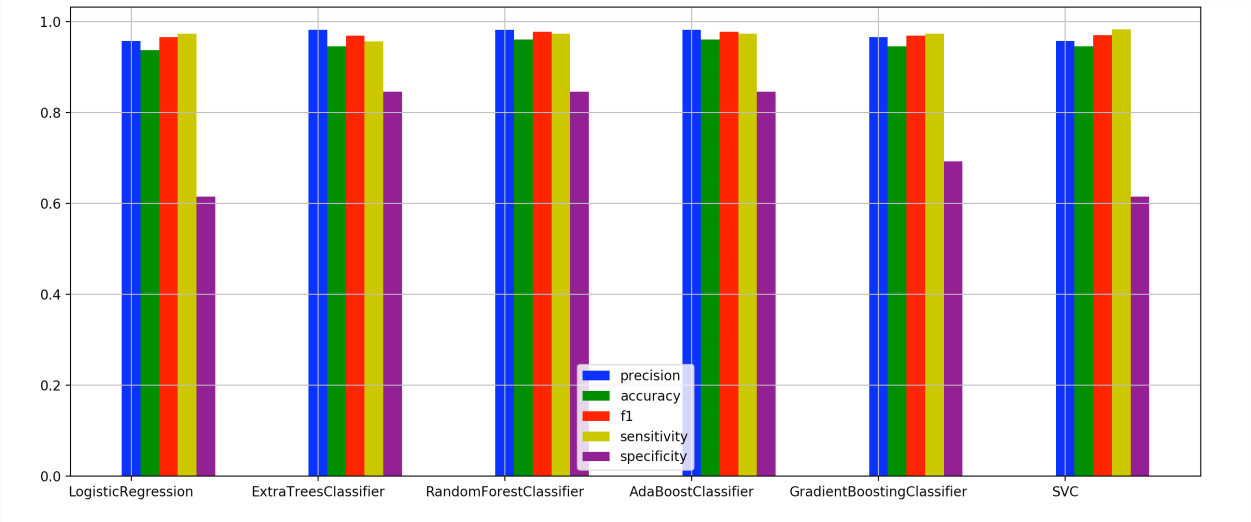


Fig.1 Performance Evaluation for Different ML models

Please explain the evaluation metrics above. Please try a different model other than the models used in the sample code. Also plot the ROC curve for the model applied.

Below are some good reference papers for your project.

Reference:

[1] Hyeongmin Kim & Yong-Min Kim ,"Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types," *Scientific Reports,* **volume 8**, Article number: 6041 (2018) ,https://www.nature.com/articles/s41598-018-24379-y

[2] Marieke Lydia Kuijjer, Joseph Nathaniel Paulson, Peter Salzman, Wei Ding & John Quackenbush, "Cancer subtype identification using somatic mutation data,", British Journal of Cancervolume 118, pages1492–1501 (2018).