A Project Report on

"Clustering of Cancerous Profiles Using Machine Learning"

Submitted in partial fulfillment of the requirement for VI semester
Bachelor of Technology
In
COMPUTER SCIENCE & ENGINEERING
REVA University.

Ruchitra K.
ID: R16CS349

TABLE OF CONTENTS

CHAPTER-01

ABSTRACT

There are various treatment options available for cancer. The type of treatment recommended for an individual depends on factors such as cancer type, severity (stage), and genetic heterogeneity. In such a complex environment, targeted drug treatments may be ineffective or yield varying responses. To study anticancer drug responses, it is essential to understand cancerous profiles. These profiles contain information that can reveal the underlying factors responsible for cancer growth. Therefore, analyzing cancer data is crucial for predicting optimal treatment options. Such analysis can help identify potential drug targets and drugs. This paper aims to provide a machine learning-based classification technique for cancerous profiles.

CHAPTER 2

INTRODUCTION

All living organisms are composed of cells, the basic unit of life. Individual cells exhibit complex functionalities. Genetics studies inherited features and how they are passed from generation to generation. It also examines gene expression levels to determine the up regulation or down regulation of genes. Researchers and scientists are working to uncover hidden aspects and networks that can aid in the proper diagnosis and treatment of diseases like cancer. Data mining and machine learning approaches are powerful tools for such data-driven analyses. Gene expression involves retrieving information from genes to synthesize functional products like proteins. The amount of m-RNA produced by a gene at a given time corresponds to its expression level. Cancer patients often show heterogeneous drug responses, with only a small subset responding to a given anticancer drug. Machine learning algorithms facilitate adaptive learning, helping predict anticancer drug treatments and classify cancer patients.

CHAPTER 3

LITERATURE SURVEY

1. *"Skin Clustering Classification Using K-Means Clustering"*
   - Detects skin cancer, providing the best chance for early diagnosis.
   - Developed a system for automatic classification of skin cancer.
   - The Gabor extraction method is the most effective among all methods.

2. *"The Beneficial Techniques in Pre processing Step of Skin Cancer Detection System Comparing"*
   - Highlights the importance of pre processing in skin cancer detection.
   - Provides a good starting point for researchers in automatic skin cancer detection.
   - Useful for researchers working on skin cancer detection systems.

3. *"A Comparative Study of Various Techniques Used for Melanoma Detection"*
   - Describes different segmentation and classification methods.
   - The pre processing stage removes artifacts associated with the session.
   - Various methods for segmentation and classification are discussed.

4. *"Melanoma Detection Through K-Means Segmentation and Feature Extraction"*
   - Clustering algorithms are useful for identifying features that differentiate melanoma.
   - K-means clustering is discussed, highlighting its advantages and limitations.

CHAPTER 4

OBJECTIVE

The objectives of this work are:

- To design and develop a method for removing instances with missing values.
- To design and develop a discretization method.
- To design and develop a feature subset selection method.

CHAPTER 5

MOTIVATION

The use of high-throughput screening techniques has generated vast amounts of data, but its full utility is limited by the tools available for processing and analysis. Machine learning holds great potential for deciphering these data in the context of cancer classification and biomarker identification. However, current machine learning tools require manual processing of raw data from various sequencing platforms, which is both tedious and time-consuming. Existing classification tools lack flexibility in choosing the best feature selection algorithms and the ability to compare various learning algorithms.

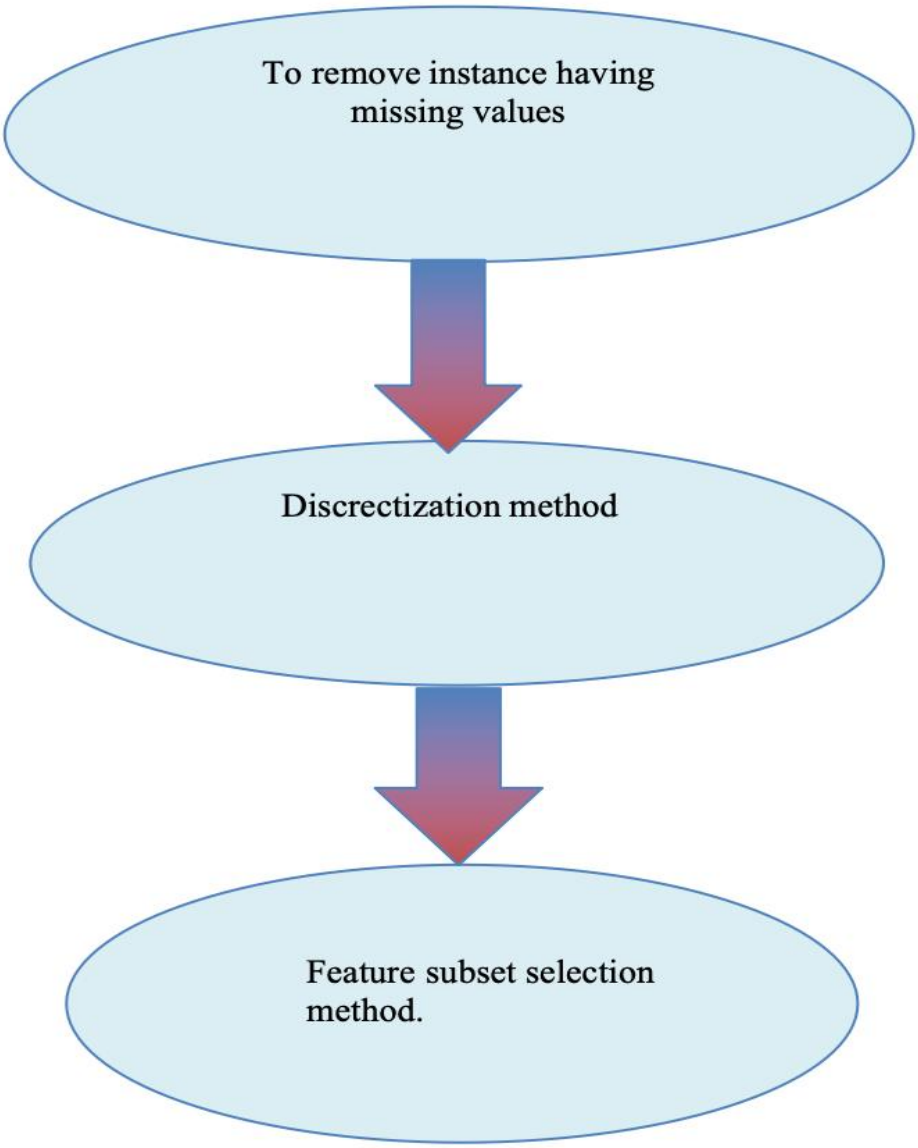
CHAPTER 6

PROBLEM STATEMENT

This project aims to explore machine learning techniques for predicting the stages of skin cancer using medical datasets. It involves data preprocessing, feature selection, and applying classification algorithms like Decision Trees, Random Forest, SVM, or Deep Learning models. The goal is to train models on labeled datasets, evaluate their performance using accuracy metrics, and visualize predictions through charts or heatmaps. The findings will help in early diagnosis and stage classification, supporting medical professionals in decision-making.

CHAPTER 7

BLOCK DIAGRAM



CHAPTER 8

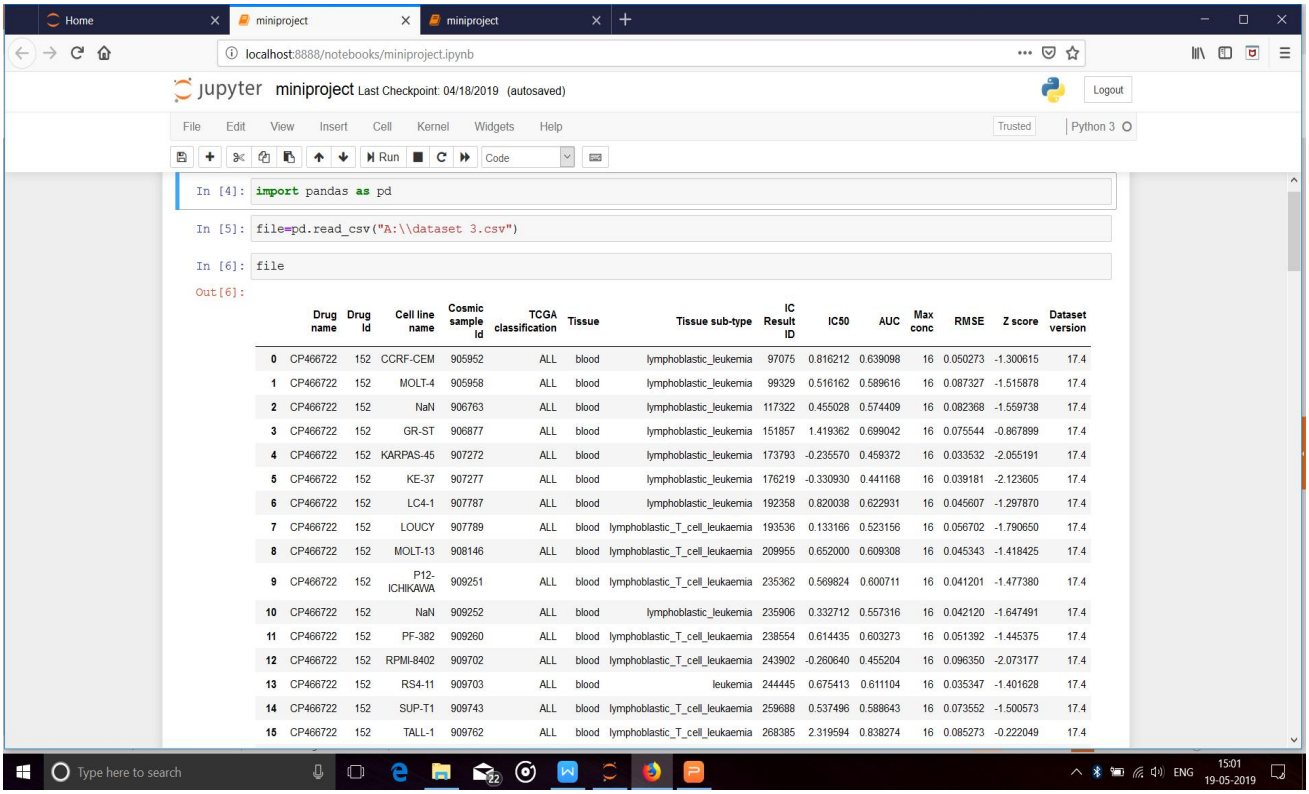
SYSTEM REQUIREMENTS

Software:

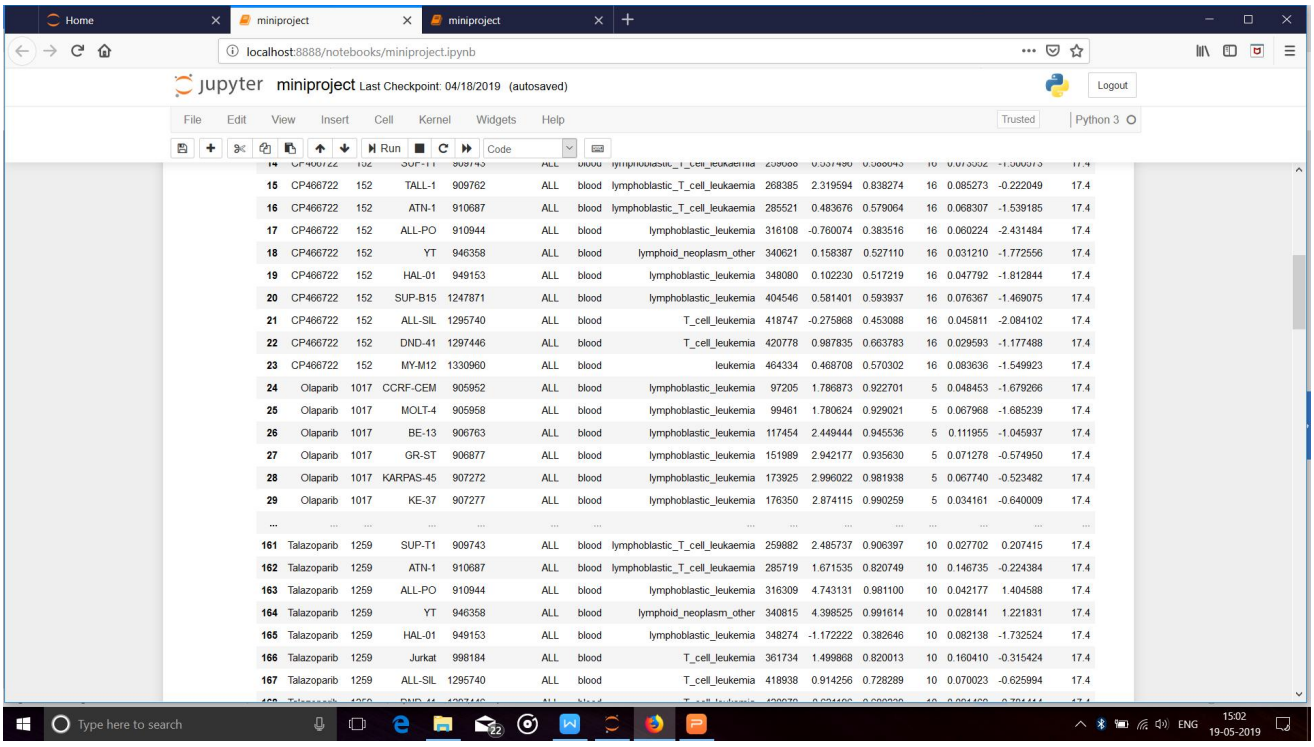
1. Anaconda-Python 3.6
2. Natural Language Toolkit (NLTK)

CHAPTER 9

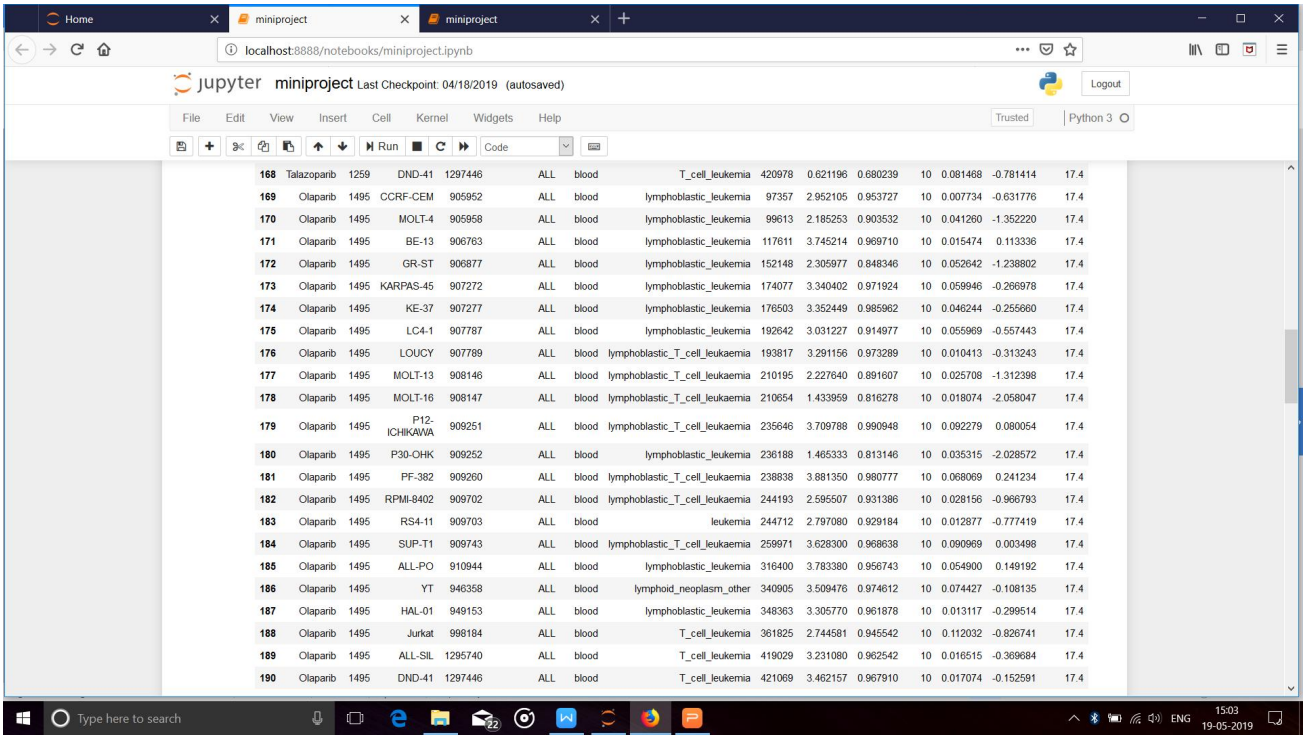
IMPLEMENTATION



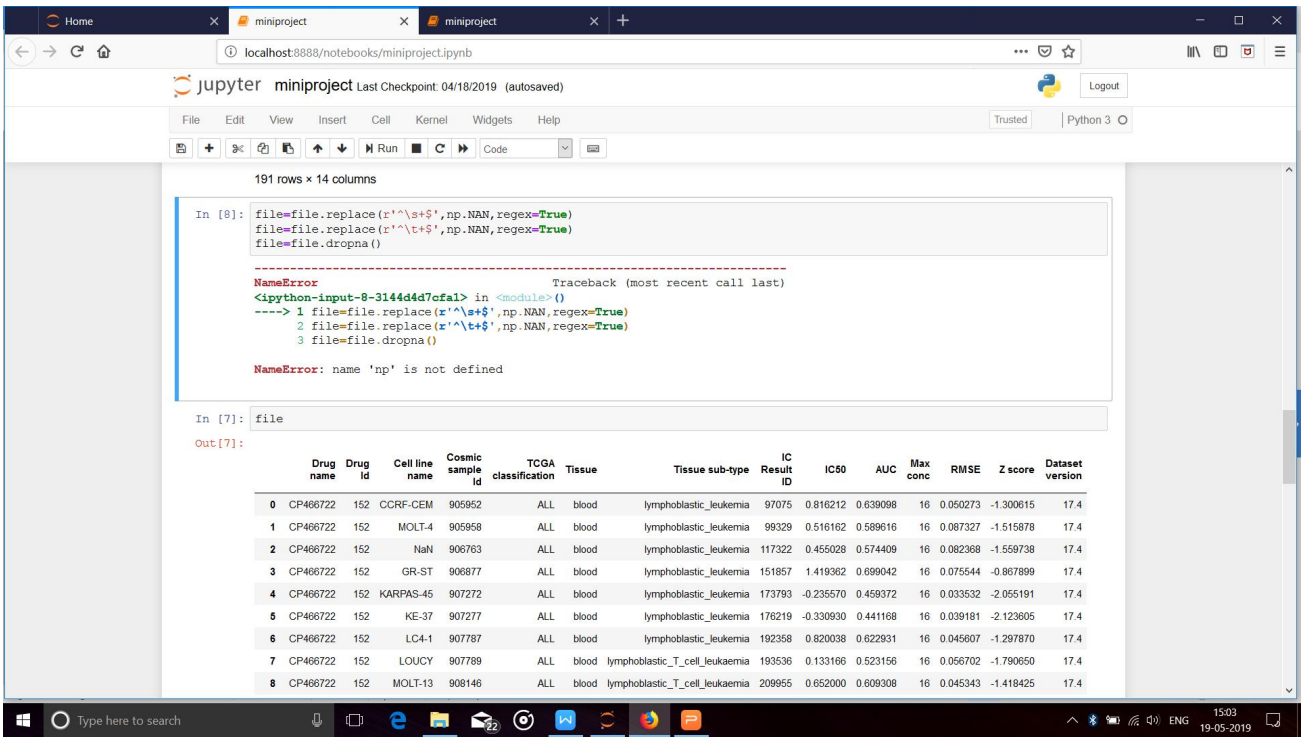
	Drug name	Drug Id	Cell line name	Cosmic sample Id	TCGA classification	Tissue	Tissue sub-type	IC Result ID	IC50	AUC	Max conc	RMSE	Z score	Dataset version
0	CP466722	152	CCRF-CEM	905952	ALL	blood	lymphoblastic_leukemia	97075	0.816212	0.639098	16	0.050273	-1.300615	17.4
1	CP466722	152	MOLT-4	905958	ALL	blood	lymphoblastic_leukemia	99329	0.516162	0.589616	16	0.087327	-1.515878	17.4
2	CP466722	152	NaN	906763	ALL	blood	lymphoblastic_leukemia	117322	0.455028	0.574409	16	0.082368	-1.559738	17.4
3	CP466722	152	GR-ST	906877	ALL	blood	lymphoblastic_leukemia	151857	1.419362	0.699042	16	0.075544	-0.867899	17.4
4	CP466722	152	KARPAS-45	907272	ALL	blood	lymphoblastic_leukemia	173793	-0.235570	0.459372	16	0.033532	-2.055191	17.4
5	CP466722	152	KE-37	907277	ALL	blood	lymphoblastic_leukemia	176219	-0.330930	0.441168	16	0.039181	-2.123605	17.4
6	CP466722	152	LC4-1	907787	ALL	blood	lymphoblastic_leukemia	192358	0.820038	0.622931	16	0.045607	-1.297870	17.4
7	CP466722	152	LOUCY	907789	ALL	blood	lymphoblastic_T_cell_leukaemia	193536	0.133166	0.523156	16	0.056702	-1.790650	17.4
8	CP466722	152	MOLT-13	908146	ALL	blood	lymphoblastic_T_cell_leukaemia	209955	0.652000	0.609308	16	0.045343	-1.418425	17.4
9	CP466722	152	P12-ICHIKAWA	909251	ALL	blood	lymphoblastic_T_cell_leukaemia	235362	0.569824	0.600711	16	0.041201	-1.477380	17.4
10	CP466722	152	NaN	909252	ALL	blood	lymphoblastic_leukemia	235906	0.332712	0.557316	16	0.042120	-1.647491	17.4
11	CP466722	152	PF-382	909260	ALL	blood	lymphoblastic_T_cell_leukaemia	238554	0.614435	0.603273	16	0.051392	-1.445375	17.4
12	CP466722	152	RPMI-8402	909702	ALL	blood	lymphoblastic_T_cell_leukaemia	243902	-0.260640	0.455204	16	0.096350	-2.073177	17.4
13	CP466722	152	RS4-11	909703	ALL	blood	leukemia	244445	0.675413	0.611104	16	0.035347	-1.401628	17.4
14	CP466722	152	SUP-T1	909743	ALL	blood	lymphoblastic_T_cell_leukaemia	259688	0.537496	0.588643	16	0.073552	-1.500573	17.4
15	CP466722	152	TALL-1	909762	ALL	blood	lymphoblastic_T_cell_leukaemia	268385	2.319594	0.838274	16	0.085273	-0.222049	17.4



15	CP466722	152	TALL-1	909762	ALL	blood	lymphoblastic_T_cell_leukaemia	268385	2.319594	0.838274	16	0.085273	-0.222049	17.4
16	CP466722	152	ATN-1	910687	ALL	blood	lymphoblastic_T_cell_leukaemia	285521	0.483676	0.579064	16	0.068307	-1.539185	17.4
17	CP466722	152	ALL-PO	910944	ALL	blood	lymphoblastic_leukemia	316108	-0.760074	0.383516	16	0.060224	-2.431484	17.4
18	CP466722	152	YT	946358	ALL	blood	lymphoid_neoplasm_other	340621	0.158387	0.527110	16	0.031210	-1.772556	17.4
19	CP466722	152	HAL-01	949153	ALL	blood	lymphoblastic_leukemia	348080	0.102230	0.517219	16	0.047792	-1.812844	17.4
20	CP466722	152	SUP-B15	1247871	ALL	blood	lymphoblastic_leukemia	404546	0.581401	0.593937	16	0.076367	-1.469075	17.4
21	CP466722	152	ALL-SIL	1295740	ALL	blood	T_cell_leukemia	418747	-0.275868	0.453088	16	0.045811	-2.084102	17.4
22	CP466722	152	DND-41	1297446	ALL	blood	T_cell_leukemia	420778	0.987835	0.663783	16	0.029593	-1.177488	17.4
23	CP466722	152	MY-M12	1330960	ALL	blood	leukemia	464334	0.468708	0.570302	16	0.083636	-1.549923	17.4
24	Olaparib	1017	CCRF-CEM	905952	ALL	blood	lymphoblastic_leukemia	97205	1.786873	0.922701	5	0.048453	-1.679266	17.4
25	Olaparib	1017	MOLT-4	905958	ALL	blood	lymphoblastic_leukemia	99461	1.780624	0.929021	5	0.067968	-1.685239	17.4
26	Olaparib	1017	BE-13	906763	ALL	blood	lymphoblastic_leukemia	117454	2.449444	0.945536	5	0.111955	-1.045937	17.4
27	Olaparib	1017	GR-ST	906877	ALL	blood	lymphoblastic_leukemia	151989	2.942177	0.935630	5	0.071278	-0.574950	17.4
28	Olaparib	1017	KARPAS-45	907272	ALL	blood	lymphoblastic_leukemia	173925	2.996022	0.981938	5	0.067740	-0.523482	17.4
29	Olaparib	1017	KE-37	907277	ALL	blood	lymphoblastic_leukemia	176350	2.874115	0.990259	5	0.034161	-0.640009	17.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
161	Talazoparib	1259	SUP-T1	909743	ALL	blood	lymphoblastic_T_cell_leukaemia	259882	2.485737	0.906397	10	0.027702	0.207415	17.4
162	Talazoparib	1259	ATN-1	910687	ALL	blood	lymphoblastic_T_cell_leukaemia	285719	1.671535	0.820749	10	0.146735	-0.224384	17.4
163	Talazoparib	1259	ALL-PO	910944	ALL	blood	lymphoblastic_leukemia	316309	4.743131	0.981100	10	0.042177	1.404588	17.4
164	Talazoparib	1259	YT	946358	ALL	blood	lymphoid_neoplasm_other	340815	4.398525	0.991614	10	0.028141	1.221831	17.4
165	Talazoparib	1259	HAL-01	949153	ALL	blood	lymphoblastic_leukemia	348274	-1.172222	0.382646	10	0.082138	-1.732524	17.4
166	Talazoparib	1259	Jurkat	998184	ALL	blood	T_cell_leukemia	361734	1.499868	0.820013	10	0.160410	-0.315424	17.4
167	Talazoparib	1259	ALL-SIL	1295740	ALL	blood	T_cell_leukemia	418938	0.914256	0.728289	10	0.070023	-0.625994	17.4



168	Talazoparib	1259	DND-41	1297446	ALL	blood	T_cell_leukemia	420978	0.621196	0.680239	10	0.081468	-0.781414	17.4
169	Olaparib	1495	CCRF-CEM	905952	ALL	blood	lymphoblastic_leukemia	97357	2.952105	0.953727	10	0.007734	-0.631776	17.4
170	Olaparib	1495	MOLT-4	905958	ALL	blood	lymphoblastic_leukemia	99613	2.185253	0.903532	10	0.041260	-1.352220	17.4
171	Olaparib	1495	BE-13	906763	ALL	blood	lymphoblastic_leukemia	117611	3.745214	0.969710	10	0.015474	0.113336	17.4
172	Olaparib	1495	GR-ST	906877	ALL	blood	lymphoblastic_leukemia	152148	2.305977	0.848346	10	0.052642	-1.238802	17.4
173	Olaparib	1495	KARPAS-45	907272	ALL	blood	lymphoblastic_leukemia	174077	3.340402	0.971924	10	0.059946	-0.266978	17.4
174	Olaparib	1495	KE-37	907277	ALL	blood	lymphoblastic_leukemia	176503	3.352449	0.985962	10	0.046244	-0.255660	17.4
175	Olaparib	1495	LC4-1	907787	ALL	blood	lymphoblastic_leukemia	192642	3.031227	0.914977	10	0.055969	-0.557443	17.4
176	Olaparib	1495	LOUCY	907789	ALL	blood	lymphoblastic_T_cell_leukaemia	193817	3.291156	0.973289	10	0.010413	-0.313243	17.4
177	Olaparib	1495	MOLT-13	908146	ALL	blood	lymphoblastic_T_cell_leukaemia	210195	2.227640	0.891607	10	0.025708	-1.312398	17.4
178	Olaparib	1495	MOLT-16	908147	ALL	blood	lymphoblastic_T_cell_leukaemia	210654	1.433959	0.816278	10	0.018074	-2.058047	17.4
179	Olaparib	1495	P12-ICHIKAWA	909251	ALL	blood	lymphoblastic_T_cell_leukaemia	235646	3.709788	0.990948	10	0.092279	0.080054	17.4
180	Olaparib	1495	P30-OHK	909252	ALL	blood	lymphoblastic_leukemia	236188	1.465333	0.813146	10	0.035315	-2.028572	17.4
181	Olaparib	1495	PF-382	909260	ALL	blood	lymphoblastic_T_cell_leukaemia	238838	3.881350	0.980777	10	0.068069	0.241234	17.4
182	Olaparib	1495	RPMI-8402	909702	ALL	blood	lymphoblastic_T_cell_leukaemia	244193	2.595507	0.931386	10	0.028156	-0.966793	17.4
183	Olaparib	1495	RS4-11	909703	ALL	blood	leukemia	244712	2.797080	0.929184	10	0.012877	-0.777419	17.4
184	Olaparib	1495	SUP-T1	909743	ALL	blood	lymphoblastic_T_cell_leukaemia	259971	3.628300	0.968638	10	0.090969	0.003498	17.4
185	Olaparib	1495	ALL-PO	910944	ALL	blood	lymphoblastic_leukemia	316400	3.783380	0.956743	10	0.054900	0.149192	17.4
186	Olaparib	1495	YT	946358	ALL	blood	lymphoid_neoplasm_other	340905	3.509476	0.974612	10	0.074427	-0.108135	17.4
187	Olaparib	1495	HAL-01	949153	ALL	blood	lymphoblastic_leukemia	348363	3.305770	0.961878	10	0.013117	-0.299514	17.4
188	Olaparib	1495	Jurkat	998184	ALL	blood	T_cell_leukemia	361825	2.744581	0.945542	10	0.112032	-0.826741	17.4
189	Olaparib	1495	ALL-SIL	1295740	ALL	blood	T_cell_leukemia	419029	3.231080	0.962542	10	0.016515	-0.369684	17.4
190	Olaparib	1495	DND-41	1297446	ALL	blood	T_cell_leukemia	421069	3.462157	0.967910	10	0.017074	-0.152591	17.4



191 rows × 14 columns

```
In [8]: file=file.replace(r'^\s+$',np.NAN,regex=True)
        file=file.replace(r'^\t+$',np.NAN,regex=True)
        file=file.dropna()
```

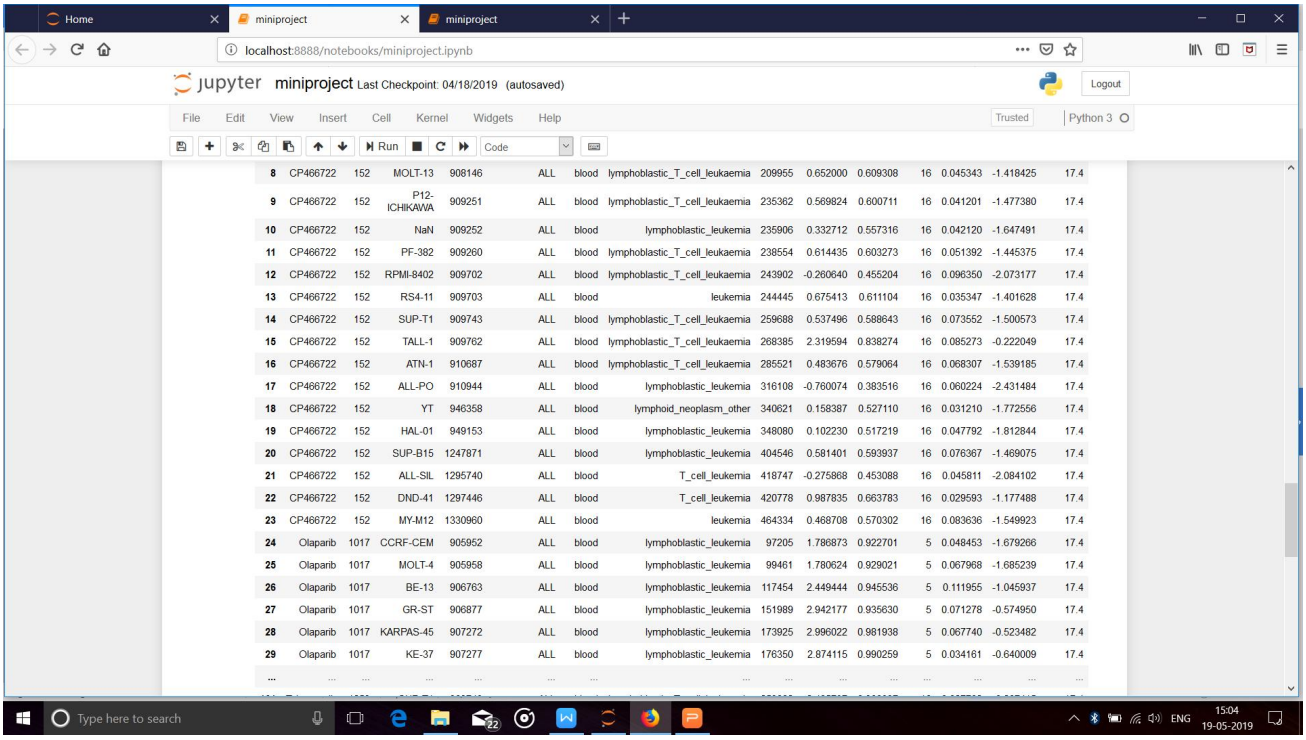
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-8-3144d4d7cfa1> in <module>()
----> 1 file=file.replace(r'^\s+$',np.NAN,regex=True)
      2 file=file.replace(r'^\t+$',np.NAN,regex=True)
      3 file=file.dropna()

NameError: name 'np' is not defined
```

```
In [7]: file
```

Out[7]:

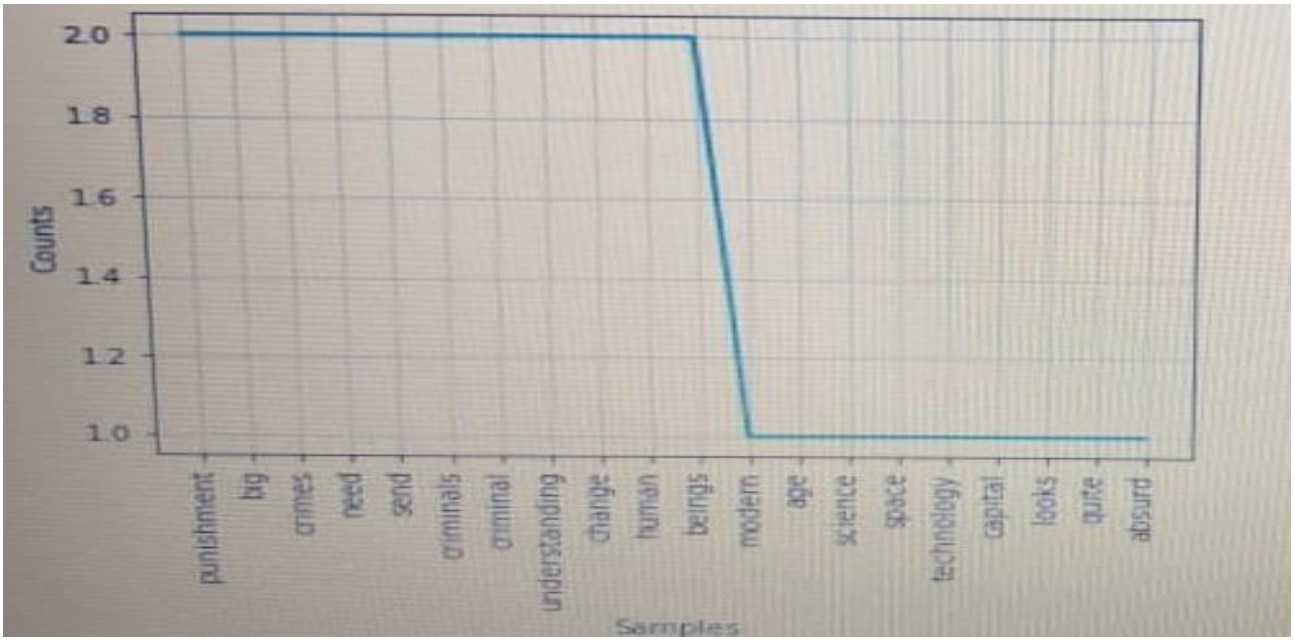
	Drug name	Drug Id	Cell line name	Cosmic sample Id	TCGA classification	Tissue	Tissue sub-type	IC Result ID	IC50	AUC	Max conc	RMSE	Z score	Dataset version
0	CP466722	152	CCRF-CEM	905952	ALL	blood	lymphoblastic_leukemia	97075	0.816212	0.639098	16	0.050273	-1.300615	17.4
1	CP466722	152	MOLT-4	905958	ALL	blood	lymphoblastic_leukemia	99329	0.516162	0.589616	16	0.087327	-1.515878	17.4
2	CP466722	152	NaN	906763	ALL	blood	lymphoblastic_leukemia	117322	0.455028	0.574409	16	0.082368	-1.559738	17.4
3	CP466722	152	GR-ST	906877	ALL	blood	lymphoblastic_leukemia	151857	1.419362	0.699042	16	0.075544	-0.867899	17.4
4	CP466722	152	KARPAS-45	907272	ALL	blood	lymphoblastic_leukemia	173793	-0.235570	0.459372	16	0.033532	-2.055191	17.4
5	CP466722	152	KE-37	907277	ALL	blood	lymphoblastic_leukemia	176219	-0.330930	0.441168	16	0.039181	-2.123605	17.4
6	CP466722	152	LC4-1	907787	ALL	blood	lymphoblastic_leukemia	192358	0.820038	0.622931	16	0.045607	-1.297870	17.4
7	CP466722	152	LOUCY	907789	ALL	blood	lymphoblastic_T_cell_leukaemia	193536	0.133166	0.523156	16	0.056702	-1.790650	17.4
8	CP466722	152	MOLT-13	908146	ALL	blood	lymphoblastic_T_cell_leukaemia	209955	0.652000	0.609308	16	0.045343	-1.418425	17.4



8	CP466722	152	MOLT-13	908146	ALL	blood	lymphoblastic_T_cell_leukaemia	209955	0.652000	0.609308	16	0.045343	-1.418425	17.4
9	CP466722	152	P12-ICHIKAWA	909251	ALL	blood	lymphoblastic_T_cell_leukaemia	235362	0.569824	0.600711	16	0.041201	-1.477380	17.4
10	CP466722	152	NaN	909252	ALL	blood	lymphoblastic_leukemia	235906	0.332712	0.557316	16	0.042120	-1.647491	17.4
11	CP466722	152	PF-382	909260	ALL	blood	lymphoblastic_T_cell_leukaemia	238554	0.614435	0.603273	16	0.051392	-1.445375	17.4
12	CP466722	152	RPMI-8402	909702	ALL	blood	lymphoblastic_T_cell_leukaemia	243902	-0.260640	0.455204	16	0.096350	-2.073177	17.4
13	CP466722	152	RS4-11	909703	ALL	blood	leukemia	244445	0.675413	0.611104	16	0.035347	-1.401628	17.4
14	CP466722	152	SUP-T1	909743	ALL	blood	lymphoblastic_T_cell_leukaemia	259688	0.537496	0.588643	16	0.073552	-1.500573	17.4
15	CP466722	152	TALL-1	909762	ALL	blood	lymphoblastic_T_cell_leukaemia	268385	2.319594	0.838274	16	0.085273	-0.222049	17.4
16	CP466722	152	ATN-1	910687	ALL	blood	lymphoblastic_T_cell_leukaemia	285521	0.483676	0.579064	16	0.068307	-1.539185	17.4
17	CP466722	152	ALL-PO	910944	ALL	blood	lymphoblastic_leukemia	316108	-0.760074	0.383516	16	0.060224	-2.431484	17.4
18	CP466722	152	YT	946358	ALL	blood	lymphoid_neoplasm_other	340621	0.158387	0.527110	16	0.031210	-1.772556	17.4
19	CP466722	152	HAL-01	949153	ALL	blood	lymphoblastic_leukemia	348080	0.102230	0.517219	16	0.047792	-1.812844	17.4
20	CP466722	152	SUP-B15	1247871	ALL	blood	lymphoblastic_leukemia	404546	0.581401	0.593937	16	0.076367	-1.469075	17.4
21	CP466722	152	ALL-SIL	1295740	ALL	blood	T_cell_leukemia	418747	-0.275868	0.453088	16	0.045811	-2.084102	17.4
22	CP466722	152	DND-41	1297446	ALL	blood	T_cell_leukemia	420778	0.987835	0.663783	16	0.029593	-1.177488	17.4
23	CP466722	152	MY-M12	1330960	ALL	blood	leukemia	464334	0.468708	0.570302	16	0.083636	-1.549923	17.4
24	Olaparib	1017	CCRF-CEM	905952	ALL	blood	lymphoblastic_leukemia	97205	1.786873	0.922701	5	0.048453	-1.679266	17.4
25	Olaparib	1017	MOLT-4	905958	ALL	blood	lymphoblastic_leukemia	99461	1.780624	0.929021	5	0.067968	-1.685239	17.4
26	Olaparib	1017	BE-13	906763	ALL	blood	lymphoblastic_leukemia	117454	2.449444	0.945536	5	0.111955	-1.045937	17.4
27	Olaparib	1017	GR-ST	906877	ALL	blood	lymphoblastic_leukemia	151989	2.942177	0.935630	5	0.071278	-0.574950	17.4
28	Olaparib	1017	KARPAS-45	907272	ALL	blood	lymphoblastic_leukemia	173925	2.996022	0.981938	5	0.067740	-0.523482	17.4
29	Olaparib	1017	KE-37	907277	ALL	blood	lymphoblastic_leukemia	176350	2.874115	0.990259	5	0.034161	-0.640009	17.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

CHAPTER 10**

RESULTS



CHAPTER 11

APPLICATION

1. Skin Cancer
2. Medical Datasets

CHAPTER 12

CONCLUSION

Cancer is a heterogeneous disease with various sub-types. Various machine learning approaches have been used to predict whether a tumor is malignant or benign. Our technique utilizes Support Vector Machines (SVM) and Neural Networks (NN) as machine learning algorithms.

FUTURE ENHANCEMENT

In the future, this approach can be extended to implement an integrative framework for anticancer drug prediction.

CHAPTER 13

REFERENCES

1. Mohd Anas, Ram Kailash Gupta, Dr. Shafeeq Ahmad, "Skin Cancer Classification Using K-Means Clustering," *International Journal of Technical Research and Application*, Volume 5, Issue 1, 2017, pp. 62-65.

2. Azadeh Noori Hoshyar, Adel Al-Jumaily, Afsaneh Noori Hoshyar, "The Beneficial Techniques in Preprocessing Step of Skin Cancer Detection System Comparing," *International Conference on Robot PRIDE 2013-2014 - Medical and Rehabilitation Robotics and Instrumentation, ConfPRIDE 2013-2014, Procedia Computer Science 42*, 2014, pp. 25-31.

3. Raina Raju K., S. Swapna Kumar, "A Comparative Study of Various Techniques Used for Melanoma Detection," *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 7, Issue 11, pp. 44-49.

4. Dr. R. Pon Periyasamy, V. Gayathiri, "Melanoma Detection Through K-Means Segmentation and Feature Extraction," *International Research Journal of Engineering and Technology (IRJET)*, Volume 4, Issue 5, 2017, pp. 1301-1305.