AI Projects

Project Rules:

1- **Number of students in each team is in rage from Four to six students from the same department.**
2- **You have three days to detect the idea you are interested in (All TA`s will be available in these three days for any explanation).**
3- **The google form for register your idea and team members will available next Tuesday at12:00 AM.**
4- **Hold your idea number.**

Idea no :1

Supervisor: Dr/ aya Nasser

Support time: Monday from 10:00 am to 12:00 am

Classifying shoulder implants in X-ray images

Introduction:

Total Shoulder Arthroplasty (TSA) is a type of surgery in which the damaged ball of the shoulder is replaced with a prosthesis. Many years later, this prosthesis may be in need of servicing or replacement. In some situations, such as when the patient has changed his country of residence, the model and the manufacturer of the prosthesis may be unknown to the patient and primary doctor. Correct identification of the implant's model prior to surgery is required for selecting the correct equipment and procedure.

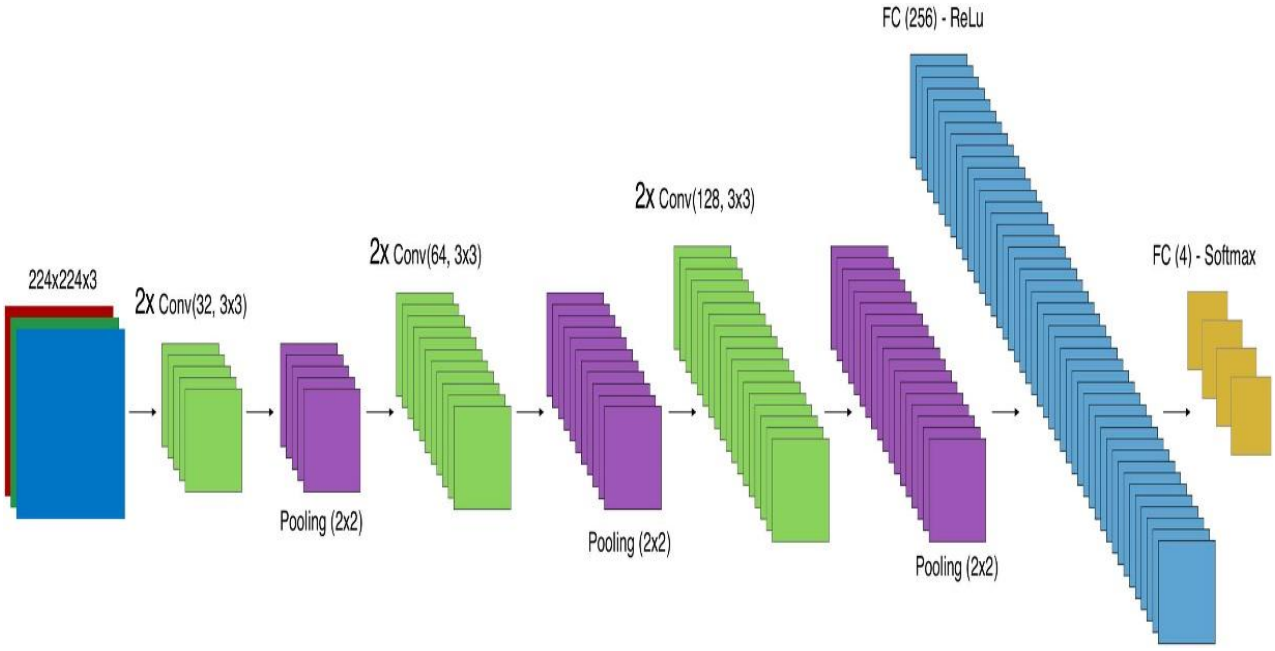
Data set description:

Images were collected by Maya Stark at BIDAL Lab at SFSU for her MS thesis project. They are from The UW Shoulder Site, manufacturer websites, and Feeley Lab at UCSF. The original collection included 605 X-ray images. Eight images that appeared to have been taken from the same patients were removed, resulting in the final 597 images. The final set contains images from the following manufacturers: 83 from Cofield, 294 from Depuy, 71 from Tornier, and 149 from Zimmer, resulting in a 4-class classification problem. Class labels are provided as the manufacturer name in file names.

Project requirements:

1- Preprocessing: 1- add black border for each image then Apply histogram normalization for each image

2- Classification: apply CNN model which represent in the following image:



Idea no :2

Supervisor: Dr/ Menna Awad

Support time: Tuesday from 4:00 pm to 6:00 pm

Loan Prediction using Machine Learning

Project idea: The idea behind this ML project is to build a model that will classify how much loan the user can take.
It is based on the user's marital status, education, number of dependents, and employment. You can build a linear model for this project.

Idea no :3

Supervisor: Dr/ Nadeen Alaa

Support time: Saturday from 3:00 pm to 4:00 pm and Monday from 10:00 am to 11:00 am

Tweets Clustering

Twitter provides a service for posting short messages. In practice, many of the tweets are very similar to each other and can be clustered together. By clustering similar tweets together, we can generate a more concise and organized representation of the raw tweets, which will be very useful for many Twitter-based applications (e.g., truth discovery, trend analysis, search ranking, etc.)

Here, the tweets are clustered using Jaccard distance metric and K-means clustering algorithm.

Jaccard Distance
The Jaccard distance, which measures dissimilarity between two sample sets (A and B). It is defined as the difference of the sizes of the union and the intersection of two sets divided by the size of the union of the sets.
Dist(A, B) = 1 - |A intersection B|/|A union B|

For example, consider the following tweets:
Tweet A: the long march
Tweet B: ides of march
|A intersection B | = 1 and |A union B | = 5, therefore the distance is 1 – (1/5)

Jaccard Distance Dist(A, B) between tweet A and B has the following properties:

1. It is small if tweet A and B are similar.
2. It is large if they are not similar.
3. It is 0 if they are the same.
4. It is 1 if they are completely different (i.e., no overlapping words).

Dataset Used
https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter

Tweets Preprocessing
Firstly, the tweets are preprocessed using the following steps:

- tweet ids and timestamps are removed
- words that starts with the symbol '@', e.g., @AnnaMedaris, are removed
- hashtag symbols are removed, e.g., #depression is converted to depression
- any URL are removed
- every word is converted to lowercase

K-Means Clustering Algorithm
K-means clustering algorithm is implemented from scratch, without using any machine learning libraries

Output results on one of the datasets is provided in "Report.pdf".

Steps To Run
5. If the system don't have python Installed in it, first install python (version greater than or equal v3.7)
    - https://www.python.org/downloads/
6. The code uses the following python native libraries - random
    - re
    - math
7. In the root directory of the project, execute the given command from command line:

- "python main.py"

8. Output from a sample run is also stored in a file named "output.txt" for the reference.

Notes :

a) The code uses "bbchealth.txt" by default for the tweets data. - A user can change the url path to another data file as desired from the given files.

b) The code uses, "3 clusters" by default and performs "5 experiments" one after another. - Each experiment increases the number of clusters being used from the previous experiment by 1. - A user can change the default value of intial clusters (k) and number of experiments to be performed.

c) The program returns the value of SSE (sum of squared error) and size of each cluster after every experiment.

d) A user can also print tweets in each cluster by uncommenting certain part of code written in the **main**.

Idea no :4

Supervisor: Dr/ Hossam

Support time: Sunday from 4:00 pm to 6:00 pm

TMDb movie data

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

● Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters.

● The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

Required

1. **Filter and clean the columns and rows** (Remove unnecessary columns & rows , Deal with NaN values with proper imputation techniques , Remove Duplicate records , apply feature scaling (normalization) for variables if necessary , Convert the used categorical columns to numerical columns using One hot encoding and label encoding techniques , check also that all columns have proper datatypes) **In order to make them tidy and be able to be fed the columns into a linear regression model.**

2. Fed the data after filtering them into a linear regression model where we will use all our selected columns as our X variables and we will use our Y variable the net profit which is the difference between (revenue_adj – budget_adj).

Note: any column that is categorical should be converted into numerical using one hot encoding or label encoding techniques , any column that has almost all values unique like id , director name , etc should be dropped from our data.

Idea no :5

Supervisor: Dr/ Alaa Tarek

Support time: Tuesday from 10:00 am to 12:00 am

Tumor Cancer Prediction

A tumor is an abnormal lump or growth of cells. When the cells in the tumor are normal, it is benign. Something just went wrong, and they overgrew and produced a lump. When the cells are abnormal and can grow uncontrollably, they are cancerous cells, and the tumor is malignant.
The early diagnosis of Tumor can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments.

The goal of the Project is to:

- Predict the Patient diagnosis based on the given features.

Dataset Snapshot

F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	F26	F27	F28	F29	F30	diagnosis
0.1175	0.2111	0.08046	324.7	0.3274	0.4228	0.4365	1.194	1.885	17.67	0.009549	1.252	0.009559	10.31	22.65	65.5	250.5	0.175	B
0.1243	0.2123	0.07254	706	0.3061	0.3407	0.5343	1.069	2.257	25.13	0.006983	0.6282	0.005617	15.2	30.15	105.3	503.2	0.1977	M
0.103	0.1662	0.06566	812.4	0.2787	0.2589	0.3542	0.6205	1.957	23.35	0.004717	0.2779	0.00313	16.57	20.86	110.3	584.1	0.1383	M
0.08574	0.1824	0.0614	2227	1.008	0.2741	0.3885	0.6999	7.561	130.2	0.003978	0.4756	0.003796	27.66	25.8	195	1482	0.2432	M
0.0641	0.159	0.05653	554.9	0.2368	0.2383	0.07061	0.8732	1.471	18.33	0.007962	0.1039	0.001906	13.46	19.76	85.67	502.5	0.05882	B
0.07834	0.1735	0.062	580.9	0.1458	0.3297	0.1958	0.905	0.9975	11.36	0.002887	0.181	0.001972	13.86	23.02	89.69	507.6	0.08388	B
0.07097	0.1516	0.06095	521.5	0.2451	0.2572	0.104	0.7655	1.742	17.86	0.006905	0.1521	0.001671	13.01	21.39	84.42	420.3	0.1099	B
0.08317	0.2035	0.06501	591.2	0.3106	0.3113	0.2658	1.51	2.59	21.57	0.007807	0.2573	0.005715	14.19	24.85	94.22	512	0.1258	B
0.09382	0.193	0.07818	185.2	0.2241	0.2932	0.1202	1.508	1.553	9.833	0.01019	0	0.0041	7.93	19.54	50.41	143.5	0	B
0.06925	0.1454	0.05549	687.6	0.2023	0.2235	0.1965	0.685	1.236	16.89	0.005969	0.1876	0.001672	14.9	23.89	95.1	537.3	0.1045	B
0.06306	0.1667	0.05474	546.7	0.2382	0.2482	0.165	0.8355	1.687	18.32	0.005996	0.1423	0.001725	13.35	19.59	86.65	463.7	0.04815	B
0.07613	0.1637	0.06343	435.9	0.1344	0.2557	0.07723	1.083	0.9812	9.332	0.0042	0.02533	0.002295	11.93	26.43	76.38	388.1	0.02832	B
0.06794	0.1592	0.05912	661.1	0.2191	0.2823	0.1072	0.6946	1.479	17.74	0.004348	0.03732	0.001802	14.67	16.93	94.17	582.7	0.05802	B
0.09031	0.1714	0.06843	701.9	0.3191	0.2849	0.2566	1.249	2.284	26.45	0.006739	0.1935	0.003747	15.11	25.63	99.43	571	0.1284	B
0.1132	0.1949	0.07292	959.5	0.7036	0.2844	0.6247	1.268	5.373	60.78	0.009407	0.6922	0.006113	17.67	29.51	119.1	645.7	0.1785	M
0.06609	0.1641	0.05764	684.5	0.1504	0.2523	0.1231	1.685	1.237	12.67	0.005371	0.0846	0.001444	14.92	25.34	96.42	609.1	0.07911	B
0.06111	0.2129	0.05025	1261	0.5506	0.4882	0.1202	1.214	3.357	54.04	0.004024	0.2249	0.001902	20.58	27.83	129.2	982	0.1185	M
0.08839	0.1848	0.06181	708.8	0.2244	0.2744	0.3167	0.895	1.804	19.36	0.00398	0.366	0.003956	15.14	25.5	101.4	575.3	0.1407	B
0.1019	0.1929	0.06744	1359	0.647	0.3187	0.3913	1.331	4.675	66.91	0.007269	0.5553	0.004232	21.2	29.41	142.1	744.7	0.2121	M
0.07421	0.1697	0.05699	1403	0.8529	0.2341	0.2117	1.849	5.632	93.54	0.01075	0.3446	0.004217	21.31	27.26	139.9	1094	0.149	M
0.06192	0.1365	0.05335	698.7	0.2244	0.2267	0.05836	0.6864	1.509	20.39	0.003338	0.01379	0.001566	14.97	16.94	95.48	566.2	0.0221	B
0.1076	0.185	0.0731	639.3	0.1931	0.4128	0.4402	0.9223	1.491	15.09	0.005251	0.3162	0.004198	14.55	29.16	99.48	529.4	0.1126	B

Dataset Description

- **Train and validation data:**

  Contains 455 row each row consist of 30 independent features(F1 -> F30) and 1 dependent feature (diagnosis)

- **Test data:**

  Contains 114 row each row consist of 30 independent features(F1 -> F30)

In the project, you will apply the followings: -

Preprocessing:

Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Classification:

- Train at least 3 models to classify each sample into distinct classes.

Model evaluation:

- Train and evaluate your classifiers on your validation set

- Idea no :6
- Supervisor: Dr/ Eman Mahmoud
- Support time: Please contact with her on her email (aboelelamemo@gmail.com)

Arabic Tweets Classification

Use a machine learning technique to classify Arabic tweets into either: positive and
negative. Your project should execute the following requirements:
1. Preprocess the tweets like removing stop words (download it from internet),
removing tashkeel and emotions and so on.
2. Make a feature engineering.
3. You can use a sentiment lexicon or not (use any one)

4. Run the data through various classifiers and calculate the accuracy and f-
score of each one to decide which one is the best.

5. Use a 20 cross validation when you report your results.

With You all the best