

Artificial Intelligence Projects

- Registration ends: 14/4/2022.
- Registration link:
https://docs.google.com/forms/d/e/1FAIpQLScvHa7JInUaIWdyH5tEp4oQ_KAa-TpWTbeYVSrgRQ14Z2Hd4g/viewform
- Minimum number of members in team is 5 and maximum is 7
- You must deliver a detailed report for the project contains all your work (Preprocessing, algorithms used in the module and the achieved accuracy).

Note: Report will be graded

Project (1): Service cancellation predictor

Description:

Service cancellation is simply when customers leave doing business with an entity. It involves determining the possibility of customers stopping doing business with an entity. In other words, if a consumer has purchased a subscription to a particular service, we must determine the likelihood that the customer would leave or cancel the membership. It is a critical prediction for many businesses because acquiring new clients often costs more than retaining existing ones. For many businesses, the ability to predict that a particular customer is at a high risk of canceling service, while there is still time to do something about it. Whereas the company will try to offer some extra functionalities for not leaving the service.

Dataset:

Dataset link: [Service Cancellation DataSet](#)

Dataset description:

The dataset consists of 7043 rows and 21 columns, where rows represent the number of customers in the dataset and the columns represent each customer's attribute. The attributes are used to predict the service cancellation of a particular customer.

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

There are 21 columns so we will divide them into independent and dependent columns:-

- 1- Independent variables: ['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges']
 - a. Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
 - b. Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

- c. Demographic info about customers – gender, age range, and if they have partners and dependents

2- Dependent variables: ['Churn']

- a. Customers who left within the last month – the column is called Churn

Main Steps:

1. You need first to apply some preprocessing on the data to make sure that it is ready to use it. Preprocessing phase includes handling unwanted features, checking if data types of columns are correct, null values, categorical values, and data scaling
2. Choose the best technique to predict the service cancellation with a high accuracy. You must prove by code that the chosen algorithm produces a higher accuracy than some of the other algorithms. try to use logistic regression, SVM, and Decision Tree ID3.

Deliverables:

An application with a simple GUI that accepts input classification method and the accuracy of the used classification method. Also, the gui can accept data of a customer and predict if he\she may cancel the subscription.

Service Cancellation Predictor

Methodology

☐ Logistic Regression ☐ SVM ☐ ID3

Train Test

Customer Data

CustomerID Partner Phone Service Online Security Tech Support Contract Monthly Charges

Gender Dependent Multiple lines Online Backup Streaming TV Paperless Billing Total Charges

Senior Citizen Tenure Internet Service Device Protection Streaming Movies Payment Method

Predict

Hint:

For GUI, you can use the **tkinter** library.

Mentor: T.A. Aya Saad

Email: aya.saad@cis.asu.edu.eg

Project (2): Bankruptcy Prediction

Bankruptcy prediction is an important problem in finance, since successful predictions would allow stakeholders to take early actions to limit their economic losses. Recently, AI models have increasingly been used in bankruptcy prediction. For any bank or financial institution, Bankruptcy prediction is of utmost importance. The aim is, therefore, to predict bankruptcy of financial institutions using machine learning classifiers.

Main steps:

Preprocessing:

- Before building your models, you need to make sure that the dataset is clean and ready-to-use (Fill empty cells , Modify data in wrong format, Deal with NaN values with proper imputation techniques , Remove Duplicate records , apply feature scaling (normalization) for variables if necessary)
- Solve imbalanced datasets problem in the given dataset using oversampling technique.

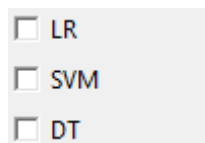
Classification:

- Train dataset using **Logistic Regression, SVM** and **Decision tree models** and print the **model accuracy** and **Confusion matrix**.

GUI:

- Design a simple graphical user interface includes:

Input: Classification method.



☒ LR
☐ SVM
☐ DT

Output: Accuracy of the selected classification model.

Dataset: [Bankruptcy Prediction Dataset](#)

Mentor: TA. Samar Aly

Email: samar.aly@cis.asu.edu.eg

Project (3): 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 or polynomial 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.

Dataset: [TMDb Movie Dataset](#)

Mentor: TA. Hazem yousef

Email: hazemyousef15@gmail.com

Project (4): 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 (Explanation)

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.

$$\text{Dist}(A, B) = 1 - |A \cap B| / |A \cup B|$$

For example, consider the following tweets:

Tweet A: the long march

Tweet B: ides of march

$|A \cap B| = 1$ and $|A \cup B| = 5$, therefore the distance is $1 - (1/5)$

Jaccard Distance $\text{Dist}(A, B)$ 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:

<https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter>

Main steps:

Tweets Preprocessing:

- 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.

- 1) 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.
- 2) The code uses, "3 clusters" by default and performs "5 experiments" one after another.
- 3) user can change the default value of initial clusters (k) and number of experiments to be performed.
- 4) The program returns the value of SSE (sum of squared error) and size of each cluster after **every** experiment (**plotted**).

Mentor:

TA: Hesham Fathy

Email: hesham.fathy@cis.asu.edu.eg

Project (5): 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)

Requirements:

In the project, you will apply the followings: -

- 1- **Preprocessing:** Before building your models, you need to make sure that the dataset is clean and ready-to-use.
- 2- **Classification:** Train at least 3 models to classify each sample into distinct classes.
- 3- **Model evaluation:** Train and evaluate your classifiers on your data set.
- 4- The user must be able to insert an input to the application, and the application has to classify that input.
- 5- After classifying the inputted sample by the three independently trained classifiers, the voting module combines their outputs to assign the inputted sample to the most frequent output (class).

Dataset: [Tumor Cancer Dataset](#)

Mentor:

TA: Alaa Tarek

Email: alaa.tarek@cis.asu.edu.eg

Project (6): Diabetes Health Indicators

Diabetes mellitus is one of the most serious chronic *illnesses* in the world. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy.

The goal of the project is to:

- Predict if the user has prediabetes/diabetes or no diabetes from the given features.

Main steps:

Preprocessing:

- Data Collection.
- Data Cleaning/ Cleansing.

Feature Selection and Extraction:

- Using any feature selection methods.
- (optional) Visualize correlation.

Model Training:

- Train using **Logistic Regression, SVM** and **Decision Tree** models.

Model Evaluation:

- Accuracy
- Precision
- F1 score
- Confusion matrix

Test the models with new data

Dataset Description:

- diabetes _ binary _ health _ indicators _ BRFSS2015.csv

dataset of 253,680 survey responses to the CDC's BRFSS2015.

- The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes.

This dataset has 21 feature variables and is not balanced.

Dataset: Diabetes Health Indicators

Mentor:

TA. Mohamed Mostafa

Email: mohamed.ahmed16121997@gmail.com

Project (7): Loan Prediction

Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant is able to repay the loan with no difficulties.

The goal of the project is to:

- Predict if the user can take loan or not from the given features.

Main steps:

Preprocessing:

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

Classification:

- Train using **Logistic Regression, SVM** and **Decision tree (ID3) models** and print the **model accuracy**.

Dataset Description:

• loan_data:

Contains 514 row each row consist of 12 independent features (Loan_ID -> Property_Area) and 1 dependent feature (Loan_Status)

Dataset: [Loan Prediction Dataset](#)

Mentor:

TA. Verena Nashaat

Email: verena.nashat98@gmail.com