EAST WEST UNIVERSITY

CSE303: Statistics for Data Science
[Spring 2023]

Project Report

Course Code : CSE303
Course Title : Statistics for Data Science
Section : 02
Group Number : 09

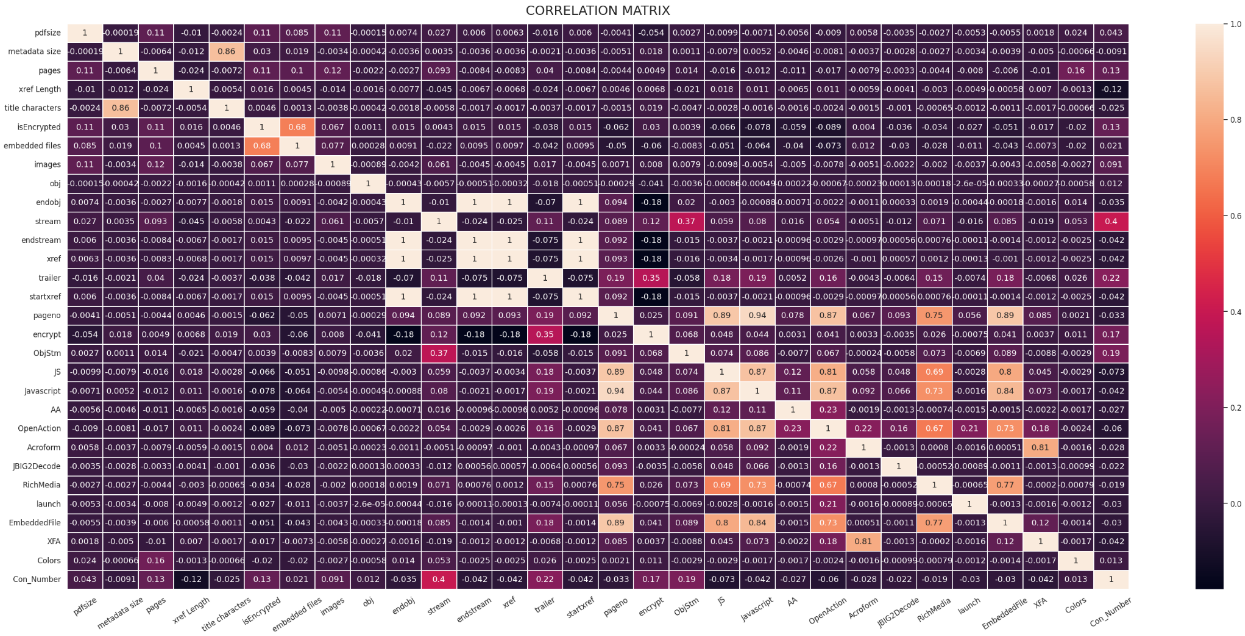
Submitted by:

Student ID	Student Name	Contribution Percentage
2020-1-60-184	**Saiful Islam**	**90%**
2020-1-60-186	**Sajib Khan**	**90%**
2020-1-60-273	**Eva Islam**	**85%**
2020-1-60-211	**Zarin Tasnim Nuzhat**	**85%**

1. Introduction

The Canadian Institute for Cybersecurity's PDF Malwear Dataset, which was acquired, contains details on PDFs that were attacked maliciously in 2022.The dataset can be used to investigate the variables that affected the prevalence of malicious acts among PDF files and to create prediction models that can calculate the likelihood of malicious acts or malicious assaults based on PDF properties.The 10023 rows and 33 columns of the dataset are provided in CSV format. The columns include pdf size, metadata size, pages, xref Length, title characters, isEncrypted, embedded files, images, obj, endobj, stream, endstream, xref, trailer, startxref, pageno, encrypt, objstm, js, javascript, AA, OpenAction, Acroform, jBig2Decode, RichMedia, launch, EmbeddedFile, XFA, colors, con_number and whether or not the PDF is under attack by any malware.To manage missing values and outliers, some cleaning of the dataset was necessary.The dataset only contains data on a portion of the PDF files that were attacked by Malware, therefore any analysis or modeling should account for the possibility of selection bias. Additionally, there weren't many missing values or duplicate values in the dataset.

2. Exploratory Data Analysis



	pdfsize	metadata size	pages	xref Length	title characters	isEncrypted	embedded files	images	obj	endobj	stream	endstream	xref	trailer	startxref	pageno	encrypt	ObjStm	JS	Javascript	AA	OpenAction	Acroform	JBIG2Decode	RichMedia	launch	EmbeddedFile	XFA	Colors	Con_Number
pdfsize	1	-0.00019	0.11	-0.01	-0.0024	0.11	0.085	0.11	-0.00015	0.0074	0.027	0.006	0.0063	-0.016	0.006	-0.0041	-0.054	0.0027	-0.0099	-0.0071	-0.0056	-0.009	0.0058	-0.0035	-0.0027	-0.0053	-0.0055	0.0018	0.024	0.043
metadata size	-0.00019	1	-0.0064	-0.012	0.86	0.03	0.019	-0.0034	-0.00042	-0.0036	0.0035	-0.0036	-0.0036	-0.0021	-0.0036	-0.0051	0.018	0.0011	-0.0079	0.0052	-0.0046	-0.0081	-0.0037	-0.0028	-0.0027	-0.0034	-0.0039	-0.005	-0.00066	-0.0091
pages	0.11	-0.0064	1	-0.024	-0.0072	0.11	0.1	0.12	-0.0022	-0.0027	0.093	-0.0084	-0.0083	0.04	-0.0084	-0.0044	0.0049	0.014	-0.016	-0.012	-0.011	-0.017	-0.0079	-0.0033	-0.0044	-0.008	-0.006	-0.01	0.16	0.13
xref Length	-0.01	-0.012	-0.024	1	-0.0054	0.016	0.0045	-0.014	-0.0016	-0.0077	-0.045	-0.0067	-0.0068	-0.024	-0.0067	0.0046	0.0068	-0.021	0.018	0.011	-0.0065	0.011	-0.0059	-0.0041	-0.003	-0.0049	-0.00058	0.007	-0.0013	-0.12
title characters	-0.0024	0.86	-0.0072	-0.0054	1	0.0046	0.0013	-0.0038	-0.00042	-0.0018	-0.0058	-0.0017	-0.0017	-0.0037	-0.0017	-0.0015	0.019	-0.0047	-0.0028	-0.0016	-0.0016	-0.0024	-0.0015	-0.001	-0.00065	-0.0012	-0.0011	-0.0017	-0.00066	-0.025
isEncrypted	0.11	0.03	0.11	0.016	0.0046	1	0.68	0.067	0.0011	0.015	0.0043	0.015	0.015	-0.038	0.015	-0.062	0.03	0.0039	-0.066	-0.078	-0.059	-0.089	0.004	-0.036	-0.034	-0.027	-0.051	-0.017	-0.02	0.13
embedded files	0.085	0.019	0.1	0.0045	0.0013	0.68	1	0.077	0.00028	0.0091	-0.022	0.0095	0.0097	-0.042	0.0095	-0.05	-0.06	-0.0083	-0.051	-0.064	-0.04	-0.073	0.012	-0.03	-0.028	-0.011	-0.043	-0.0073	-0.02	0.021
images	0.11	-0.0034	0.12	-0.014	-0.0038	0.067	0.077	1	-0.00089	-0.0042	0.061	-0.0045	-0.0045	0.017	-0.0045	0.0071	0.008	0.0079	-0.0098	-0.0054	-0.005	-0.0078	-0.0051	-0.0022	-0.002	-0.0037	-0.0043	-0.0058	-0.0027	0.091
obj	-0.00015	-0.00042	-0.0022	-0.0016	-0.00042	0.0011	0.00028	-0.00089	1	-0.00043	-0.0057	-0.00051	-0.00032	-0.018	-0.00051	-0.00029	-0.041	-0.0036	-0.00086	-0.00049	-0.00022	-0.00067	-0.00023	0.00013	0.00018	-2.6e-05	-0.00033	-0.00027	-0.00058	0.012
endobj	0.0074	-0.0036	-0.0027	-0.0077	-0.0018	0.015	0.0091	-0.0042	-0.00043	1	-0.01	1	1	-0.07	1	0.094	-0.18	0.02	-0.003	-0.00088	-0.00071	-0.0022	-0.0011	0.00033	0.0019	-0.00044	-0.00018	-0.0016	0.014	-0.035
stream	0.027	0.0035	0.093	-0.045	-0.0058	0.0043	-0.022	0.061	-0.0057	-0.01	1	-0.024	-0.025	0.11	-0.024	0.089	0.12	0.37	0.059	0.08	0.016	0.054	-0.0051	-0.012	0.071	-0.016	0.085	-0.019	0.053	0.4
endstream	0.006	-0.0036	-0.0084	-0.0067	-0.0017	0.015	0.0095	-0.0045	-0.00051	1	-0.024	1	1	-0.075	1	0.092	-0.18	-0.015	-0.0037	-0.0021	-0.00096	-0.0029	-0.00097	0.00056	0.00076	-0.00011	-0.0014	-0.0012	-0.0025	-0.042
xref	0.0063	-0.0036	-0.0083	-0.0068	-0.0017	0.015	0.0097	-0.0045	-0.00032	1	-0.025	1	1	-0.075	1	0.093	-0.18	-0.016	-0.0034	-0.0017	-0.00096	-0.0026	-0.001	0.00057	0.0012	-0.00013	-0.001	-0.0012	-0.0025	-0.042
trailer	-0.016	-0.0021	0.04	-0.024	-0.0037	-0.038	-0.042	0.017	-0.018	-0.07	0.11	-0.075	-0.075	1	-0.075	0.19	0.35	-0.058	0.18	0.19	0.0052	0.16	-0.0043	-0.0064	0.15	-0.0074	0.18	-0.0068	0.026	0.22
startxref	0.006	-0.0036	-0.0084	-0.0067	-0.0017	0.015	0.0095	-0.0045	-0.00051	1	-0.024	1	1	-0.075	1	0.092	-0.18	-0.015	-0.0037	-0.0021	-0.00096	-0.0029	-0.00097	0.00056	0.00076	-0.00011	-0.0014	-0.0012	-0.0025	-0.042
pageno	-0.0041	-0.0051	-0.0044	0.0046	-0.0015	-0.062	-0.05	0.0071	-0.00029	0.094	0.089	0.092	0.093	0.19	0.092	1	0.025	0.091	0.89	0.94	0.078	0.87	0.067	0.093	0.75	0.056	0.89	0.085	0.0021	-0.033
encrypt	-0.054	0.018	0.0049	0.0068	0.019	0.03	-0.06	0.008	-0.041	-0.18	0.12	-0.18	-0.18	0.35	-0.18	0.025	1	0.068	0.048	0.044	0.0031	0.041	0.0033	-0.0035	0.026	-0.00075	0.041	0.0037	0.011	0.17
ObjStm	0.0027	0.0011	0.014	-0.021	-0.0047	0.0039	-0.0083	0.0079	-0.0036	0.02	0.37	-0.015	-0.016	-0.058	-0.015	0.091	0.068	1	0.074	0.086	-0.0077	0.067	-0.00024	-0.0058	0.073	-0.0069	0.089	-0.0088	-0.0029	0.19
JS	-0.0099	-0.0079	-0.016	0.018	-0.0028	-0.066	-0.051	-0.0098	-0.00086	-0.003	0.059	-0.0037	-0.0034	0.18	-0.0037	0.89	0.048	0.074	1	0.87	0.12	0.81	0.058	0.048	0.69	-0.0028	0.8	0.045	-0.0029	-0.073
Javascript	-0.0071	0.0052	-0.012	0.011	-0.0016	-0.078	-0.064	-0.0054	-0.00049	-0.00088	0.08	-0.0021	-0.0017	0.19	-0.0021	0.94	0.044	0.086	0.87	1	0.11	0.87	0.092	0.066	0.73	-0.0016	0.84	0.073	-0.0017	-0.042
AA	-0.0056	-0.0046	-0.011	-0.0065	-0.0016	-0.059	-0.04	-0.005	-0.00022	-0.00071	0.016	-0.00096	-0.00096	0.0052	-0.00096	0.078	0.0031	-0.0077	0.12	0.11	1	0.23	-0.0019	-0.0013	-0.00074	-0.0015	-0.0015	-0.0022	-0.0017	-0.027
OpenAction	-0.009	-0.0081	-0.017	0.011	-0.0024	-0.089	-0.073	-0.0078	-0.00067	-0.0022	0.054	-0.0029	-0.0026	0.16	-0.0029	0.87	0.041	0.067	0.81	0.87	0.23	1	0.22	0.16	0.67	0.21	0.73	0.18	-0.0024	-0.06
Acroform	0.0058	-0.0037	-0.0079	-0.0059	-0.0015	0.004	0.012	-0.0051	-0.00023	-0.0011	-0.0051	-0.00097	-0.001	-0.0043	-0.00097	0.067	0.0033	-0.00024	0.058	0.092	-0.0019	0.22	1	-0.0013	0.0008	-0.0016	0.00051	0.81	-0.0016	-0.028
JBIG2Decode	-0.0035	-0.0028	-0.0033	-0.0041	-0.001	-0.036	-0.03	-0.0022	0.00013	0.00033	-0.012	0.00056	0.00057	-0.0064	0.00056	0.093	-0.0035	-0.0058	0.048	0.066	-0.0013	0.16	-0.0013	1	-0.00052	-0.00089	-0.0011	-0.0013	-0.00099	-0.022
RichMedia	-0.0027	-0.0027	-0.0044	-0.003	-0.00065	-0.034	-0.028	-0.002	0.00018	0.0019	0.071	0.00076	0.0012	0.15	0.00076	0.75	0.026	0.073	0.69	0.73	-0.00074	0.67	0.0008	-0.00052	1	-0.00065	0.77	-0.0002	-0.00079	-0.019
launch	-0.0053	-0.0034	-0.008	-0.0049	-0.0012	-0.027	-0.011	-0.0037	-2.6e-05	-0.00044	-0.016	-0.00011	-0.00013	-0.0074	-0.00011	0.056	-0.00075	-0.0069	-0.0028	-0.0016	-0.0015	0.21	-0.0016	-0.00089	-0.00065	1	-0.0013	-0.0016	-0.0012	-0.03
EmbeddedFile	-0.0055	-0.0039	-0.006	-0.00058	-0.0011	-0.051	-0.043	-0.0043	-0.00033	-0.00018	0.085	-0.0014	-0.001	0.18	-0.0014	0.89	0.041	0.089	0.8	0.84	-0.0015	0.73	0.00051	-0.0011	0.77	-0.0013	1	0.12	-0.0014	-0.03
XFA	0.0018	-0.005	-0.01	0.007	-0.0017	-0.017	-0.0073	-0.0058	-0.00027	-0.0016	-0.019	-0.0012	-0.0012	-0.0068	-0.0012	0.085	0.0037	-0.0088	0.045	0.073	-0.0022	0.18	0.81	-0.0013	-0.0002	-0.0016	0.12	1	-0.0017	-0.042
Colors	0.024	-0.00066	0.16	-0.0013	-0.00066	-0.02	-0.02	-0.0027	-0.00058	0.014	0.053	-0.0025	-0.0025	0.026	-0.0025	0.0021	0.011	-0.0029	-0.0029	-0.0017	-0.0017	-0.0024	-0.0016	-0.00099	-0.00079	-0.0012	-0.0014	-0.0017	1	0.013
Con_Number	0.043	-0.0091	0.13	-0.12	-0.025	0.13	0.021	0.091	0.012	-0.035	0.4	-0.042	-0.042	0.22	-0.042	-0.033	0.17	0.19	-0.073	-0.042	-0.027	-0.06	-0.028	-0.022	-0.019	-0.03	-0.03	-0.042	0.013	1

We have create a correlation data table where,
Positive correlation range: <0.4
Weak correlation range: >0.0999

No correlation range: <=0.0999
Negative correlation range: >0

Columns	Positive correlation	Weak positive correlation	No correlation	Negetive correlation
pdfsize	1	3	11	15
metadata size	2	0	6	22
pages	1	6	4	19
xref Length	1	0	8	21
title characters	2	0	3	25
isEncrypted	2	3	13	12
embedded files	2	1	12	15
images	1	2	8	19
obj	1	0	5	24
endobj	4	0	8	18
stream	2	4	13	12
endstream	4	0	6	20
xref	4	0	6	20
trailer	1	9	4	16
startxref	4	0	6	20
pageno	6	1	15	8
encrypt	1	14	6	9

objstm	1	2	13	14
js	6	2	7	15
javascript	6	1	8	15
AA	1	3	4	22
OpenAction	6	6	4	14
Acroform	2	1	9	18
jBig2Decode	1	1	8	20
RichMedia	6	1	9	14
launch	1	1	1	27
EmbeddedFile	6	2	4	18
XFA	2	2	6	20
colors	1	1	7	21
con_number	1	6	5	18

3. Data Preprocessing

- **Data collection:** We have collected mour dataset for an online source. Our online source is Canadian Institute for Cybersecurity. From this site we have worked on the CIC-Evasive-PDFMal2022 dataset.
- **Data cleaning:** To process our dataset we have to find out the duplicate and missing data. In our dataset we have a total number of 10026 columns and 33 rows.In the dataset we did not find any duplicate data and we did have only 3 missing data; we have dropped the missing values from the data set. Then we have 10023 columns and 33 rows.
- **Data reduction:** To reduce the noise from diffirent features we replaced the appropriate data with respect to the features.In total we had 17 columns where we found noise.
- **Data storage:** We store the processed data in a spreadsheet and for our coding we have the dataset as a csv file that is easily accessible for analysis.
- **Data analysis:** We have used machine learning algorithms to identify patterns and visualization tools to identify the most common types of malware.To predict the malwear we made a model using logistic regression and linear regression.

4. Machine Learning Models

Logistic Regression: A statistical model called logistic regression is used to forecast a binary result (0 or 1) from one or more predictor variables.The link between the predictor variables and the outcome variable is modeled using a logistic function.Modeling the link between a dependent variable and one or more independent variables is done using this form of regression analysis.The logistic regression model converts a linear combination of the input data into a probability score between 0 and 1 using a logistic function, often known as a sigmoid function. The logistic function's formula is as follows:

$g(z) = 1 / (1 + e^(-z))$

The objective is to minimize the difference between the true binary outcomes and the expected probabilities, which is measured by the function.

Support Vector Machine(SVM): Support vector machine, or SVM for short, is a well-liked machine learning technique that can be applied to both classification and regression applications. The objective of SVM in classification jobs is to identify a hyperplane that can most effectively classify the data points.The margin—the separation between the hyperplane and the nearest data points from each class—is maximized by selecting the hyperplane. Support vectors are utilized to define the hyperplane; these are the points that are closest to the hyperplane.By utilizing various kernel functions, SVM can handle both linearly and non-linearly separable data. Better class separation is possible because of the kernel function's mapping of the input features into a higher-dimensional space where the data is linearly separable.

In our project we have used logistic regression and linear regression machine learning models.

- **Logistic regression:** With the help of a confusing matrix and performance measure we predicted the dataset and found out accuracy, f-score,precision and recall.

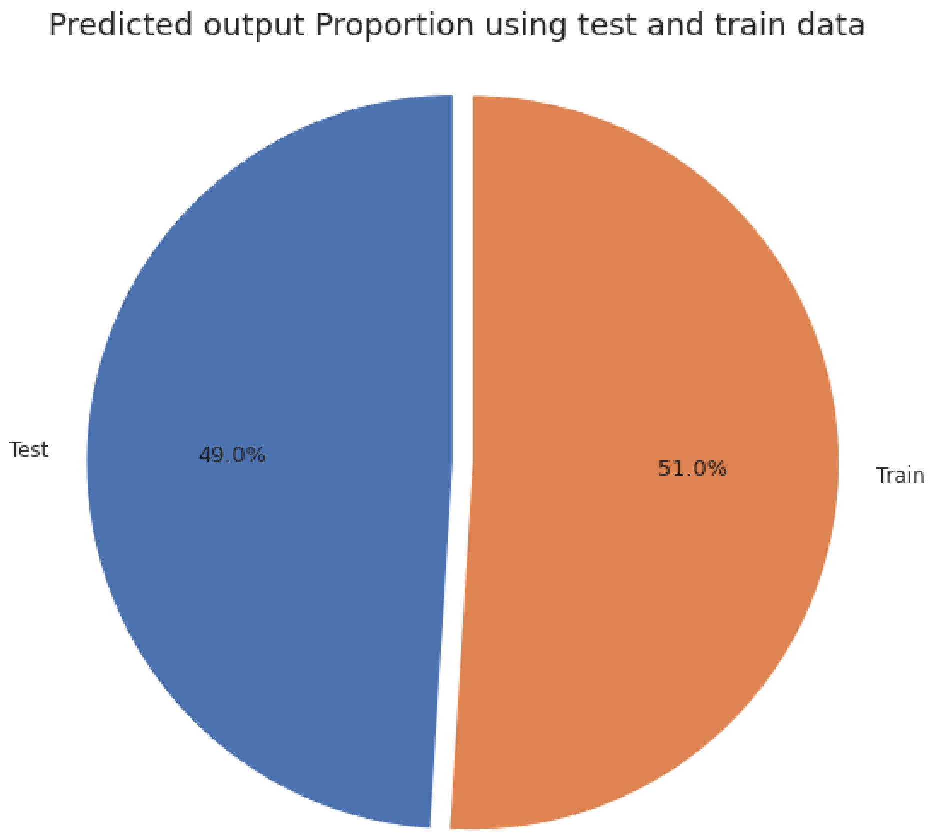
From this we can tell how perfectly the model will result in future data.

- **Linear regression:** With the help of regression we can predict the data with MAE,MSE and R2 score.

In the algorithm we have used grid search to get the best parameter from our dataset.

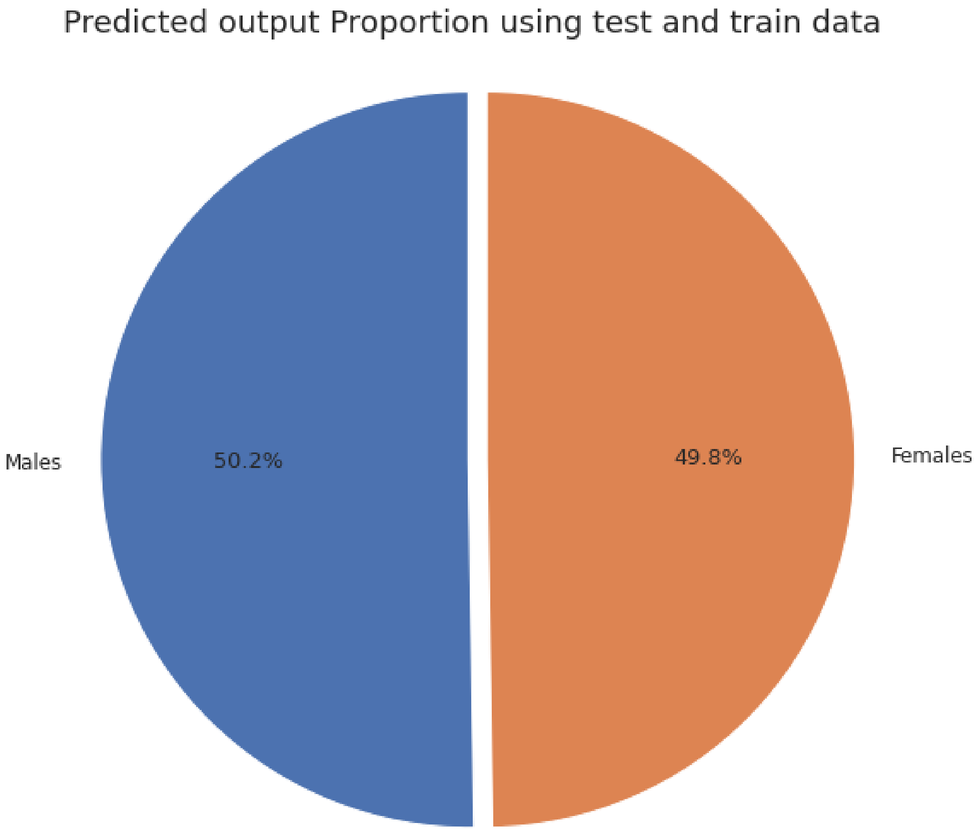
5. Performance Evaluation and Discussion

We calculated accuracy, precision, F1-score, and recall to assess the effectiveness of our model using logistic regression.90% is the score and 90% is the accuracy of the logistic regression model using default parameters. As a result, the model is too well fitted.



In the following pie chart we have found the difference between predicted output of test and train data.

To cope up with the overfit problem we have used 5 fold cross validation.



We used five fold cross validation to address this issue. After cross-validation, the mean score is 91%, while the minimum score is 90%. The outcome didn't significantly improve.90% of the model's predictions for these test and training data are favorable.