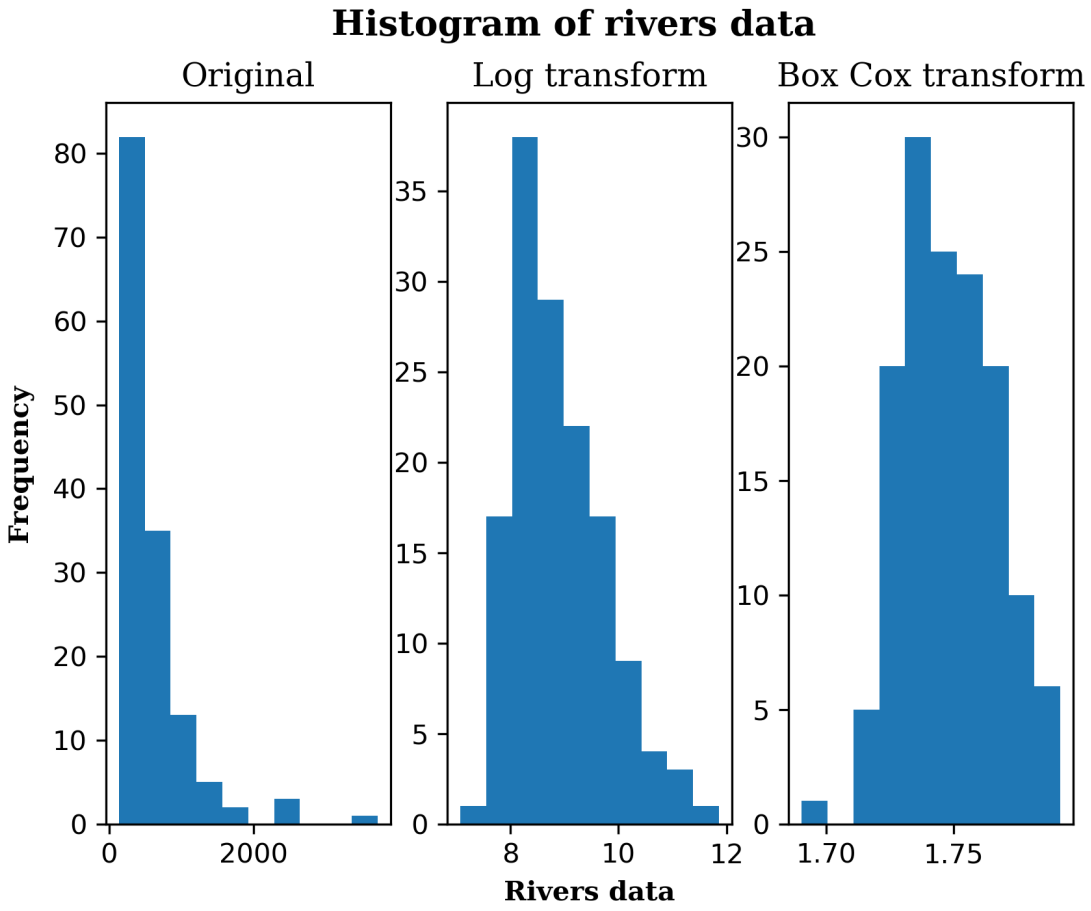
Q1



Discussion:

In the above graph the left plot is the original rivers data, the middle plot has a log2 transformation applied to the data, while the right plot has a Box Cox transformation applied to the data. In the original data the variance of the data is forced to the left of the plot making the overall distribution a right skew. Both the middle and right plots have transformations applied that force the data into a more normal distribution. I would select the Box Cox transformation to continue with data analysis, because the log transformation still has a slight right skew, while the Box Cox has a more normal distribution. Box Cox operates on the formula given below, where lambda is an input value between -5 and 5. If lambda is 0 then a simple log transformation is done, otherwise the data is transformed according to the division with lambda shown in the equation. I used the default values of scipy stats so lambda was selected for me. The scipy function selects lambda based on the value that maximises the log likelihood function of the data. Now that the data has been transformed to normal, excessive emphasis won't be places on the majority data in the machine learning models and statistical analyses.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & if \quad \lambda \neq 0 \\ \log(y) & if \quad \lambda = 0 \end{cases}$$

Q2

```
Iris data original class list:
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]

Iris data dummy encoded class list:
 [[1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0],
[1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1
, 0], [1, 0], [1, 0], [1, 0], [1, 0], [1, 0], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0,
1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1]
, [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 1], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0],
[0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0
, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0], [0, 0]]
```

Q3

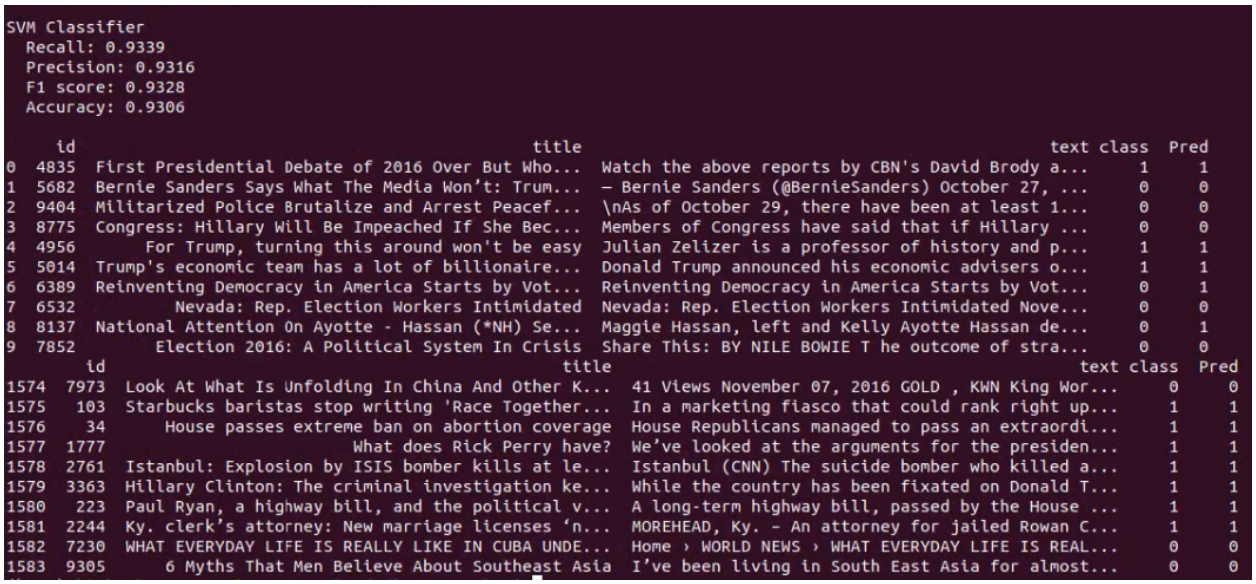
A and B: I got my dataset from Kaggle, see reference below. Since this is a static dataset which I downloaded and included with my submission I didn't see the need to write it again to a file and reupload it into the python program as per the B requirement. As you can see from the program I was effectively able to load the data into the program from a csv file and then do train test split with random seed.

C and D: Preprocessing is shown in the screenshots below, tfidf is implemented in the pipeline function.

```
# CountVectorizer includes preprocessing: lowercase, tokenize, lemmatizing, stemming, vectorization
# removes: stop words, punctuation, space, low words
text_clf = Pipeline([ ('vect', CountVectorizer(stop_words='english')), \
                      ('tfidf', TfidfTransformer()), \
                      ('clf', SGDClassifier(loss='hinge', penalty='l2', \
                                            alpha=.00001, random_state=0, \
                                            max_iter=1000, tol=0.001)) ])
```

```
['15th', '1914', '1922', '1988', '1990', '2000', '2010', '2012', 'achieved', 'adapt', 'adopt', 'affairs', 'affluence', 'agenda', 'agree', 'america', 'anchor', 'anew', 'apostate',
 'appreciation', 'apt', 'asked', 'aspired', 'astound', 'astutely', 'aurophobes', 'austan', 'bad', 'bank', 'beat', 'begins', 'believes', 'boot', 'borrowing', 'bryan', 'called', 'c
alling', 'calvinball', 'camp', 'capital', 'cares', 'caricature', 'carries', 'caused', 'ceased', 'center', 'champion', 'civil', 'club', 'coinage', 'com', 'come', 'commitment', 'co
nclusion', 'congress', 'congressional', 'conservative', 'considerable', 'continues', 'contradicted', 'correctly', 'counsel', 'create', 'creditors', 'curiously', 'currency', 'davi
d', 'deaf', 'dear', 'debate', 'debtors', 'decision', 'deferential', 'deflation', 'delight', 'deluded', 'demonstrably', 'depreciation', 'depression', 'described', 'designed', 'det
ermine', 'detroit', 'dim', 'disappointing', 'disclosure', 'distinction', 'distinguished', 'does', 'don', 'donald', 'draft', 'duty', 'economic', 'economist', 'economists', 'econom
y', 'eichengreen', 'elaborated', 'elected', 'election', 'electoral', 'elite', 'elitist', 'emphasis', 'encountered', 'encouraged', 'erudite', 'ethnic', 'exclusive', 'extraordinary
', 'faded', 'fails', 'fair', 'fallacy', 'far', 'farmers', 'favors', 'fed', 'file', 'fixation', 'forbes', 'forced', 'forgiven', 'forgotten', 'framed', 'free', 'french', 'friedman'
, 'friend', 'frum', 'frums', 'gave', 'general', 'global', 'gold', 'gone', 'good', 'goolsbee', 'governor', 'grand', 'great', 'grotesque', 'growth', 'gubernatorial', 'guru', 'happe
ned', 'having', 'head', 'high', 'highlighting', 'history', 'honorable', 'host', 'hybrid', 'ideas', 'ignore', 'immortal', 'including', 'indiana', 'induced', 'instrument', 'integri
ty', 'international', 'interview', 'interwar', 'item', 'jabberwocky', 'jack', 'jacques', 'jaws', 'jennings', 'journal', 'july', 'karen', 'kemp', 'keynesian', 'klinghard', 'krugma
n', 'labor', 'late', 'later', 'leadership', 'lee', 'left', 'let', 'like', 'listen', 'long', 'longer', 'lost', 'magnificently', 'major', 'majority', 'massive', 'mate', 'mcintosh',
 'mckinley', 'memory', 'mess', 'midst', 'mike', 'misguided', 'misguidedly', 'misunderstand', 'monetary', 'months', 'nations', 'neo', 'neurotic', 'news', 'note', 'noted', 'objecti
ve', 'observe', 'occupied', 'office', 'onset', 'operations', 'opportunity', 'org', 'pac', 'painful', 'parity', 'paul', 'pence', 'perfect', 'persuaded', 'perverse', 'picking', 'pi
nk', 'platform', 'play', 'playing', 'plovers', 'point', 'policy', 'politics', 'post', 'powerful', 'pray', 'pre', 'prescribing', 'presidency', 'president', 'presidential', 'previo
usly', 'pro', 'probably', 'proper', 'properly', 'prosperity', 'provides', 'public', 'pursue', 'qe2', 'qualities', 'quite', 'race', 'races', 'radio', 'rank', 'ravaging', 'rave', '
reached', 'really', 'reason', 'reasons', 'recalled', 'recent', 'recently', 'recognized', 'red', 'reference', 'replaced', 'replied', 'report', 'reported', 'request', 'restoration'
, 'rethink', 'return', 'reuter', 'reviled', 'right', 'robert', 'role', 'roses', 'rueff', 'rules', 'run', 'running', 'seat', 'secular', 'seeking', 'sensed', 'serve', 'served', 'se
rvice', 'shirk', 'showed', 'silver', 'similarities', 'similarity', 'slam', 'slightly', 'small', 'sound', 'soundly', 'speech', 'spending', 'splendidly', 'standard', 'started', 'st
ate', 'stated', 'street', 'stubbornly', 'success', 'super', 'superior', 'supplied', 'support', 'syndicated', 'talk', 'think', 'thinkprogress', 'tim', 'time', 'timothy', 'tone', '
track', 'true', 'trump', 'turned', 'tweet', 'twice', 'ultimately', 'unsuccessful', 'unsuccessfully', 'urged', 'usual', 'vacating', 'vice', 'victory', 'violets', 'vox', 'wall', 'w
anted', 'war', 'ways', 'went', 'wife', 'william', 'win', 'wit', 'words', 'world', 'worsening', 'writing', 'wrong', 'wrote', 'zoellick']
```

E: The screenshot below shows the performance metrics from the SVM classifier used on the fake and real news dataset. I used this dataset to predict based on the article text which articles are fake (class=0) or real (class=1). I tuned both the hyperparameters, preprocessing steps and the model selection in order to obtain the best performance scores. I initially started with a Naive Bayes classifier which gave me ~88% accuracy, I then switch to SVM which initially gave me ~90% accuracy, after tuning the hyperparameters and preprocessing I was able to obtain 93% accuracy.



```
SVM Classifier
  Recall: 0.9339
  Precision: 0.9316
  F1 score: 0.9328
  Accuracy: 0.9306
```

	id	title	text	class	Pred
0	4835	First Presidential Debate of 2016 Over But Who...	Watch the above reports by CBN's David Brody a...	1	1
1	5682	Bernie Sanders Says What The Media Won't: Trum...	— Bernie Sanders (@BernieSanders) October 27, ...	0	0
2	9404	Militarized Police Brutalize and Arrest Peacef...	\nAs of October 29, there have been at least 1...	0	0
3	8775	Congress: Hillary Will Be Impeached If She Bec...	Members of Congress have said that if Hillary ...	0	0
4	4956	For Trump, turning this around won't be easy	Julian Zelizer is a professor of history and p...	1	1
5	5014	Trump's economic team has a lot of billionaire...	Donald Trump announced his economic advisers o...	1	1
6	6389	Reinventing Democracy in America Starts by Vot...	Reinventing Democracy in America Starts by Vot...	0	1
7	6532	Nevada: Rep. Election Workers Intimidated	Nevada: Rep. Election Workers Intimidated Nove...	0	0
8	8137	National Attention On Ayotte - Hassan (*NH) Se...	Maggie Hassan, left and Kelly Ayotte Hassan de...	0	1
9	7852	Election 2016: A Political System In Crisis	Share This: BY NILE BOWIE T he outcome of stra...	0	0

	id	title	text	class	Pred
1574	7973	Look At What Is Unfolding In China And Other K...	41 Views November 07, 2016 GOLD , KWN King Wor...	0	0
1575	103	Starbucks baristas stop writing 'Race Together...	In a marketing fiasco that could rank right up...	1	1
1576	34	House passes extreme ban on abortion coverage	House Republicans managed to pass an extraordi...	1	1
1577	1777	What does Rick Perry have?	We've looked at the arguments for the presiden...	1	1
1578	2761	Istanbul: Explosion by ISIS bomber kills at le...	Istanbul (CNN) The suicide bomber who killed a...	1	1
1579	3363	Hillary Clinton: The criminal investigation ke...	While the country has been fixated on Donald T...	1	1
1580	223	Paul Ryan, a highway bill, and the political v...	A long-term highway bill, passed by the House ...	1	1
1581	2244	Ky. clerk's attorney: New marriage licenses 'n...	MOREHEAD, Ky. - An attorney for jailed Rowan C...	1	1
1582	7230	WHAT EVERYDAY LIFE IS REALLY LIKE IN CUBA UNDE...	Home › WORLD NEWS › WHAT EVERYDAY LIFE IS REAL...	0	0
1583	9305	6 Myths That Men Believe About Southeast Asia	I've been living in South East Asia for almost...	0	0

F: I feel very good about my model's performance, my accuracy is well above chance and generally anything in the 90% range i'm happy with. Precision and recall are both very high which shows the model is accurately able to detect positive values (class=1) and label them as positive. The F1 score is also high which shows the model is not skewed to doing well on positive values and badly on negative (class=0), but instead it does a great job on labelling both classes. This is a very balanced model and it highlights the all rounded good performance I generally look for in tuning models. I would say my model performance is good but not excellent, I generally reserve excellent status for models that achieve over 95% performance.

References

1) Plummer, A. (Sept. 2020). Box-Cox Transformation: Explained What the Box-Cox transformation is and how to implement it in Python. Towards Data Science: https://towardsdatascience.com/box-cox-transformation-explained-51d745e34203
2) Fake or real news. Kaggle: https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news?select=fake_or_real_news.csv