Foundation to Data Science

SARFARAZ JAMAL-17093

ASSIGNMENT 1

Problem Statement

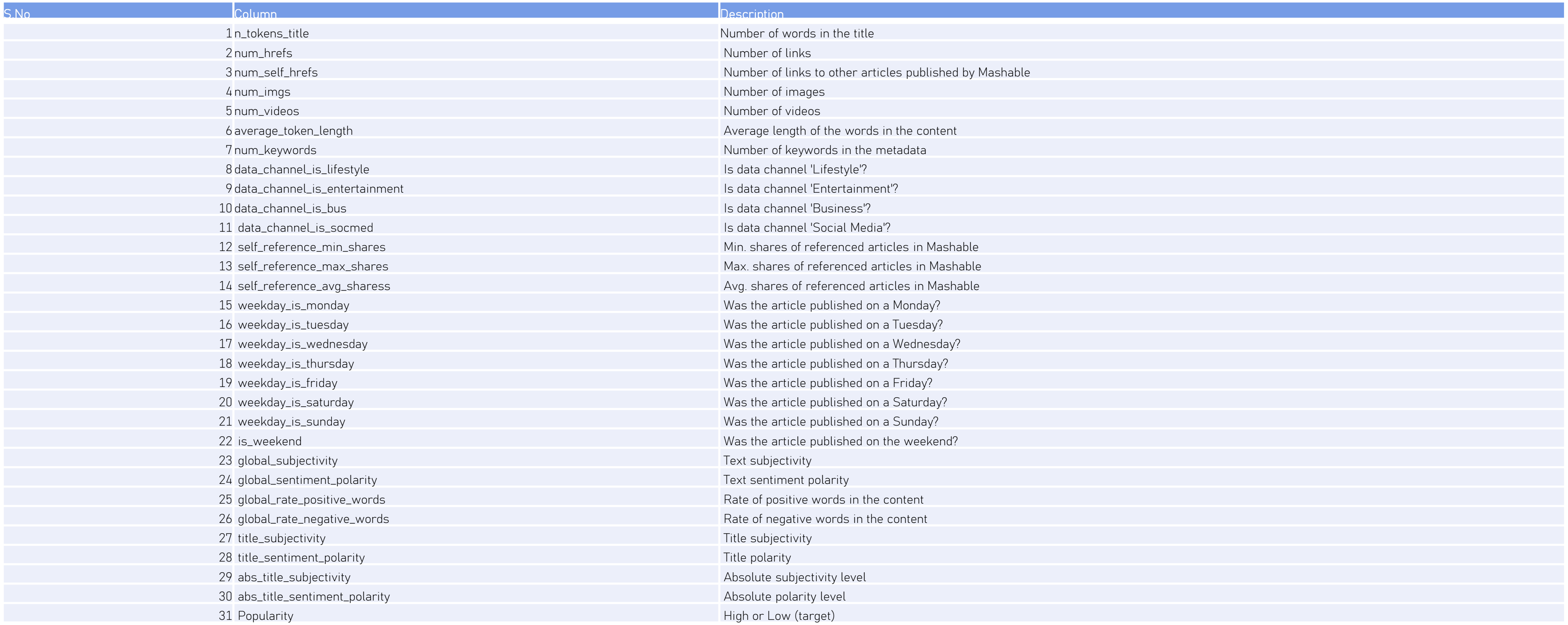
- Given a data set of 30 features, the project aims to accurately predict the popularity of news articles of the news outlet Mashable.
- It is a binary classification problem with Low Popularity and High Popularity as the two class labels.

Data Understanding

- The data provided consists of **30 explanatory variables and 1 target variable.**
- It includes **2000 samples**.

Data Understanding

Summary Table

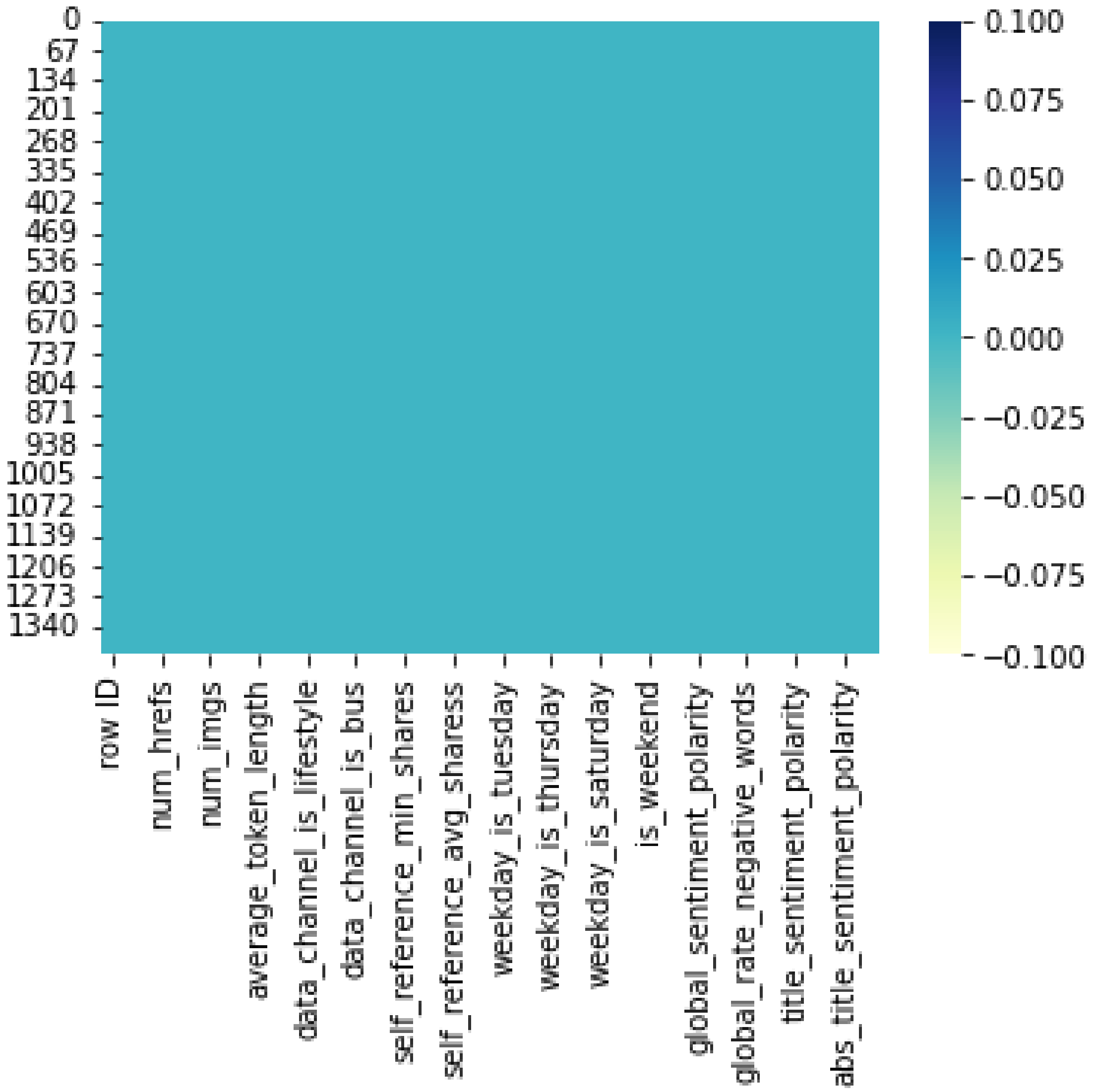


S.No	Column	Description
1	n_tokens_title	Number of words in the title
2	num_hrefs	Number of links
3	num_self_hrefs	Number of links to other articles published by Mashable
4	num_imgs	Number of images
5	num_videos	Number of videos
6	average_token_length	Average length of the words in the content
7	num_keywords	Number of keywords in the metadata
8	data_channel_is_lifestyle	Is data channel 'Lifestyle'?
9	data_channel_is_entertainment	Is data channel 'Entertainment'?
10	data_channel_is_bus	Is data channel 'Business'?
11	data_channel_is_socmed	Is data channel 'Social Media'?
12	self_reference_min_shares	Min. shares of referenced articles in Mashable
13	self_reference_max_shares	Max. shares of referenced articles in Mashable
14	self_reference_avg_sharess	Avg. shares of referenced articles in Mashable
15	weekday_is_monday	Was the article published on a Monday?
16	weekday_is_tuesday	Was the article published on a Tuesday?
17	weekday_is_wednesday	Was the article published on a Wednesday?
18	weekday_is_thursday	Was the article published on a Thursday?
19	weekday_is_friday	Was the article published on a Friday?
20	weekday_is_saturday	Was the article published on a Saturday?
21	weekday_is_sunday	Was the article published on a Sunday?
22	is_weekend	Was the article published on the weekend?
23	global_subjectivity	Text subjectivity
24	global_sentiment_polarity	Text sentiment polarity
25	global_rate_positive_words	Rate of positive words in the content
26	global_rate_negative_words	Rate of negative words in the content
27	title_subjectivity	Title subjectivity
28	title_sentiment_polarity	Title polarity
29	abs_title_subjectivity	Absolute subjectivity level
30	abs_title_sentiment_polarity	Absolute polarity level
31	Popularity	High or Low (target)

Data Understanding

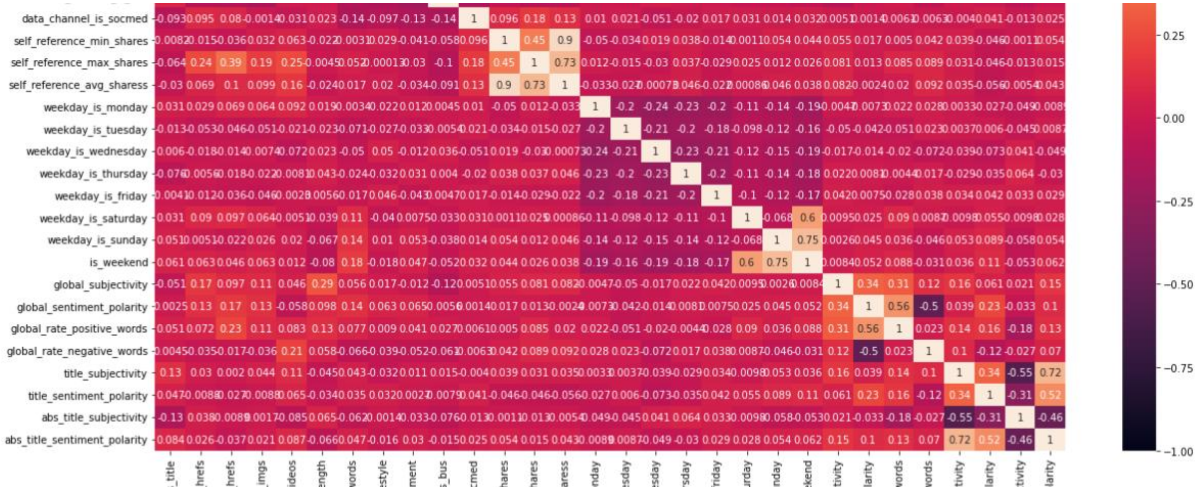
Missing Values

- No Missing Values were found in the data



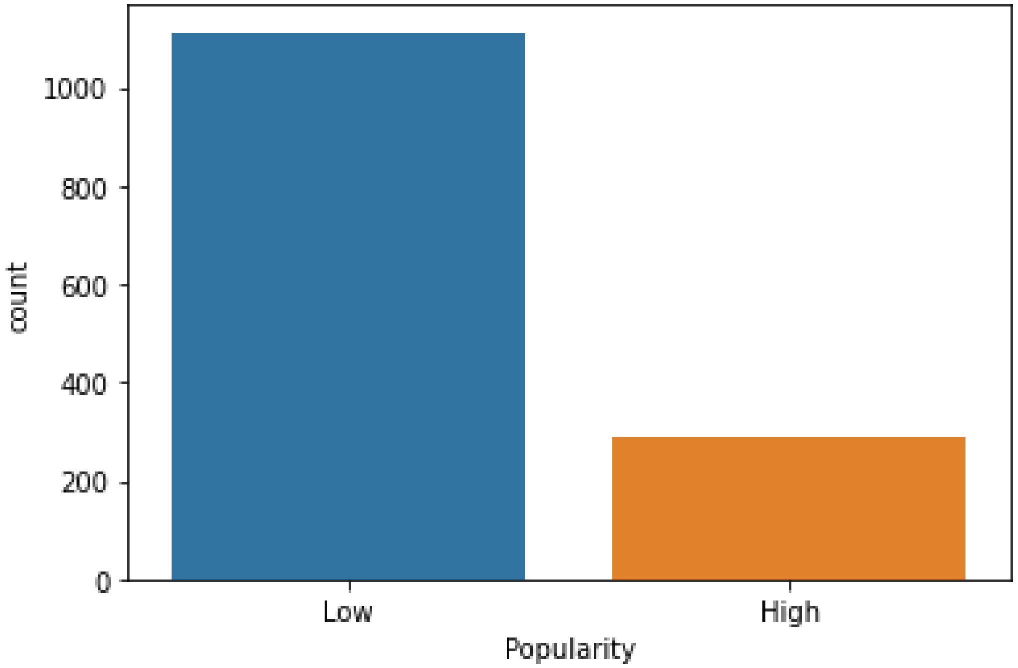
Data Understanding
Correlation Matrix

- No significant correlation was found between attributes



Data Understanding
Class Imbalance

- Significant class imbalance between class labels



Data Preparation

- The data was split into **training and testing** sets, with **1400 samples** as training data and the remaining **600 samples** as testing data.
- The training data was further split into **training and validation** sets to in order to evaluate models locally before final submission. The ratio was **90% training and 10% validation**
- The **column "row ID"** was removed. All other columns were included in model training
- The data was further divided into **predictive features and target feature (X and y)**
- The y variable was converted into a binary variable with **1 being "Low" and 0 being "High".**

Data Modelling

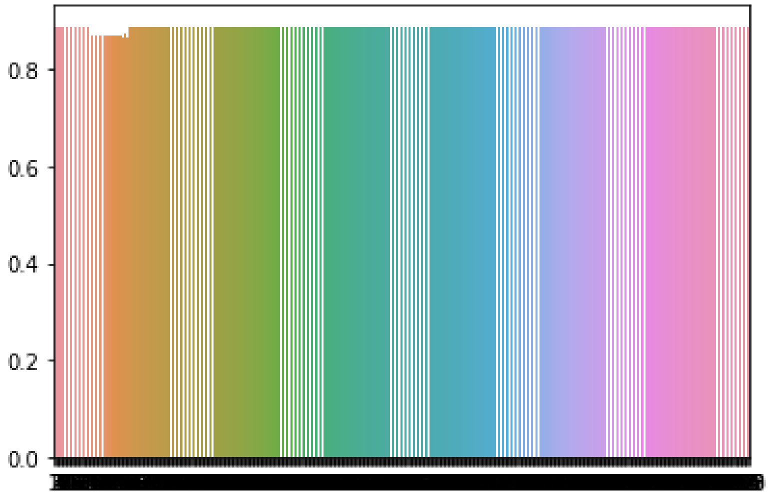
- Total 8 attempts were made and the summary is as follows:

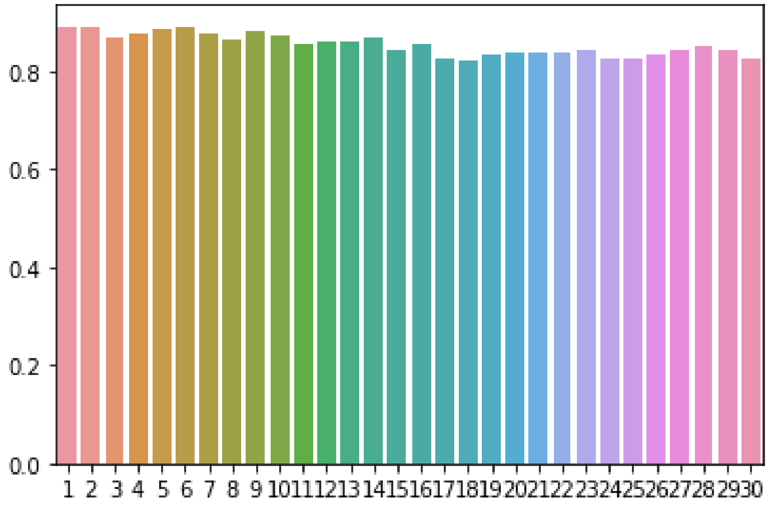
Submission No	Score	Comments
1	0.69444	Default Decision Tree
2	0.81666	max_depth = 1
3	0.81666	max_depth=2
4	0.8111	max_depth =3
5	0.81666	max_depth = 3, min_samples_leaf = 60
6	0.81666	ccp_alpha = 0.004
7	0.78333	Default KNN
8	0.81111	leaf_size =1, n_neighbors=27, p=1

Data Modelling

Decision Trees with max_depth and min_sample_leaf tuning

- Plotting different Depths and Minimum Sample Leafs against F-Scores, we find that the **optimal depth is 2** and **optimal minimum leaf size is 60**

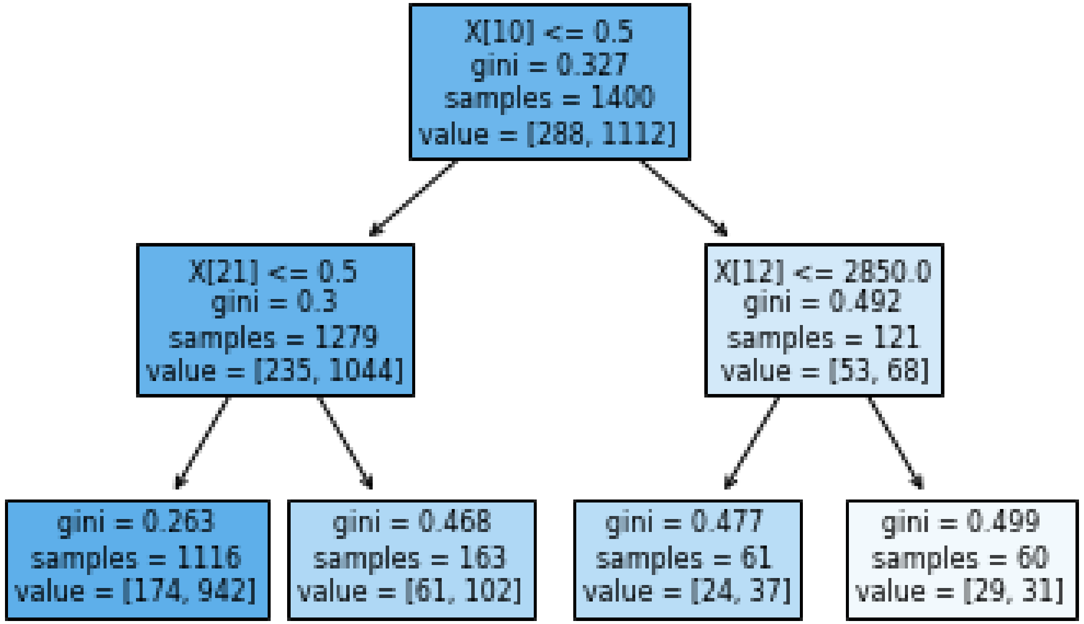




Data Modelling

Decision Trees with max_depth=2 and min_sample_leaf=60

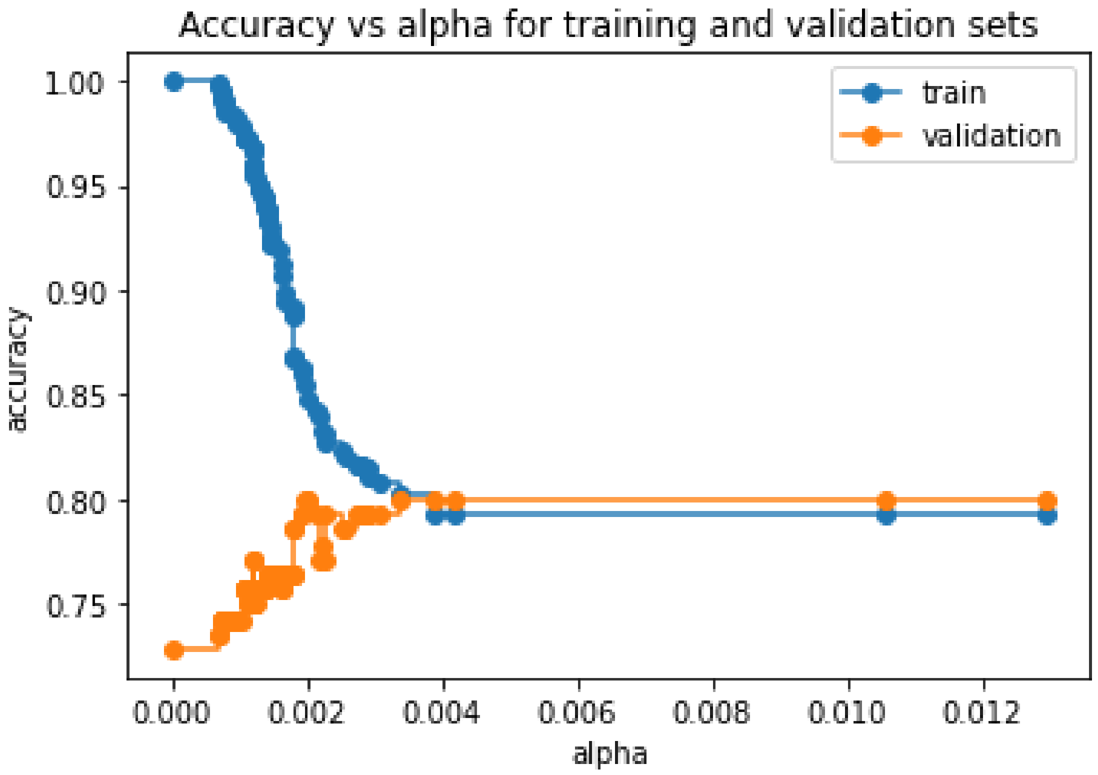
- The F-score in the final testing data was 0.81666



Data Modelling

Decision Trees with ccp_alpha tuning

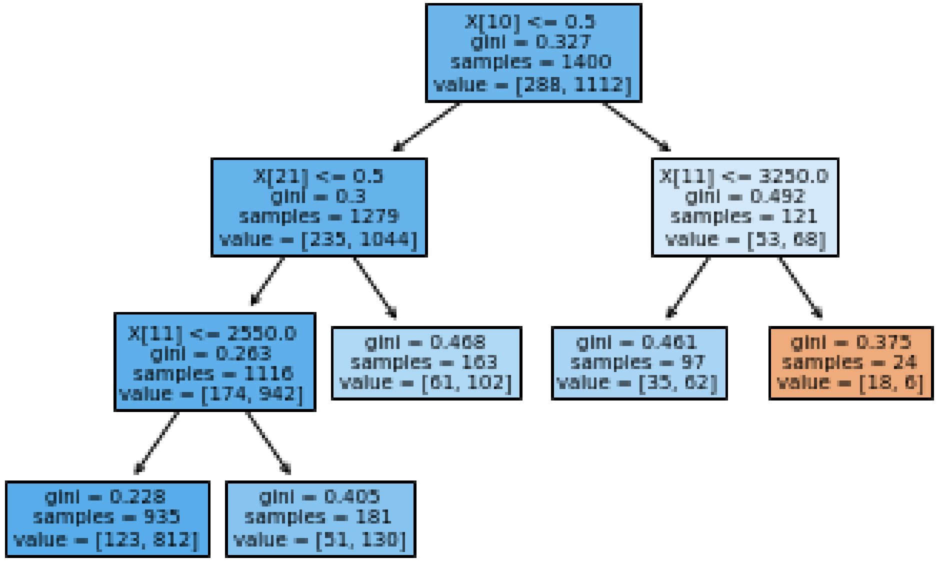
- The graph suggests the optimal value of ccp_alpha is 0.04 onwards.



Data Modelling

Decision Trees with ccp_alpha = 0.004

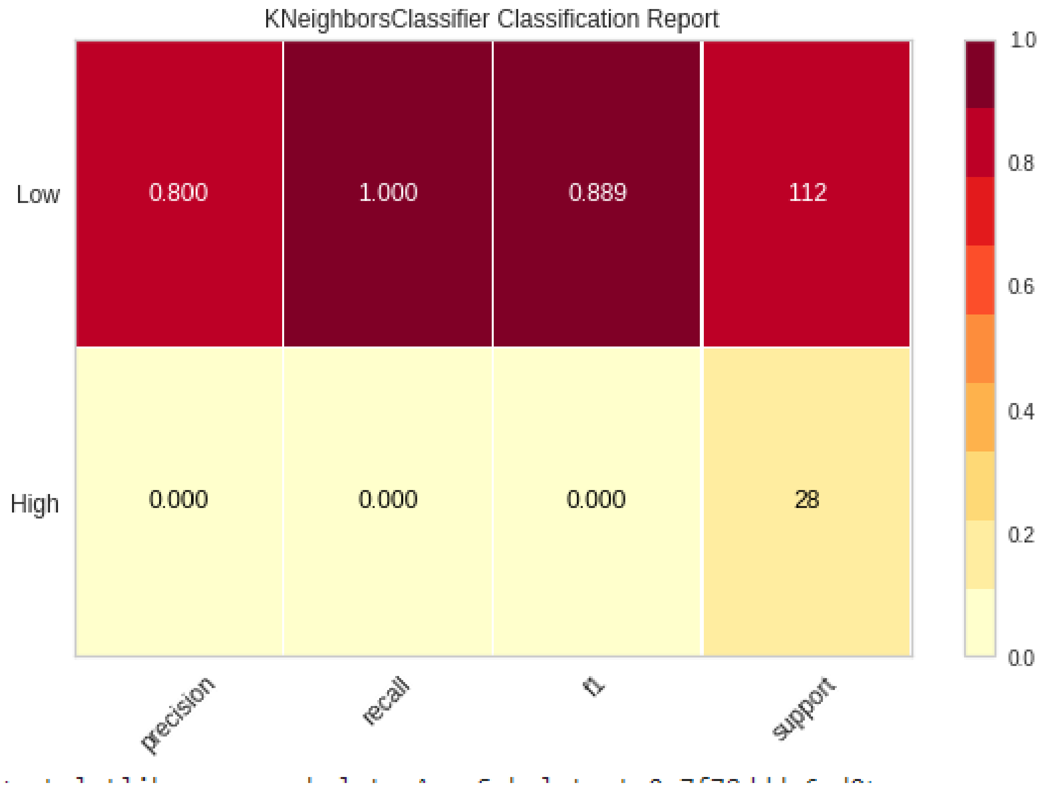
- The F-score in the final testing data was 0.81666



Data Modelling

K-nearest neighbours with leaf_size =1, n_neighbors=27, p=1

- The classification report based on validation data is as follows.
- The F-score in the final testing data was 0.81666



Links

- https://www.kaggle.com/sarfarazjamal/sarfarazjamal17093assignment1/notebook?scriptVersionId=89410172

Thank You