# Related Work

There have been various implementations of machine learning algorithms for the prediction of mobile phone price range. Many of these use the ‘Mobile Price Classification’ dataset from Kaggle. Approaches include artificial neural networks (Nasser & Al-Shawwa, 2019), decision trees and naïve bayes classifiers (Saiteja, 2019). Asim & Khan (2018) also use the same two algorithms on data collected from GSMArena.com.

The work we aim to do in this report largely mimic the aims described but with data collected ourselves. This allows for more freedom and choice in our features and permits the option of validation with results from related work. We also aim to test additional algorithms such as logistical regression.

# Data Exploration

The large variety and number of features in a mobile phone makes the prediction/classification of prices an ideal problem for machine learning. However, a drawback of this is that not all mobile phones share features found in other phones. For example, in our dataset, about 47% of the phones have selfie cameras.

Another related issue is the numerous presences of null values throughout the data. This may be because of the issue described previously – the phone does not have these features, or because of some mislabelling issue. This is shown in the figure below, where the white pixels represent null values. It is clear that some pre-processing is required.

A picture containing building

Description automatically generated

Figure : Matrix showing null values (generated using msno)

We also plot the correlation heatmap to measure nullity correlation – that is, how strongly the presence or absence of one variable affects the presence of another. Interestingly we see that if a phone has selfie\_camera\_single, then there is some correlation that it also has no main\_camera\_single.

A screenshot of a cell phone

Description automatically generated

Figure : Correlation Heatmap

After cleaning the data, we plot several boxplots of all the features grouped by their price range. 0 represents phones in the budget range less than 300$, 1 represents mid range phones between 300$ and 700$ and 3 represents the high end phones. Note that most features with zero values are imputed to represent a separate class.

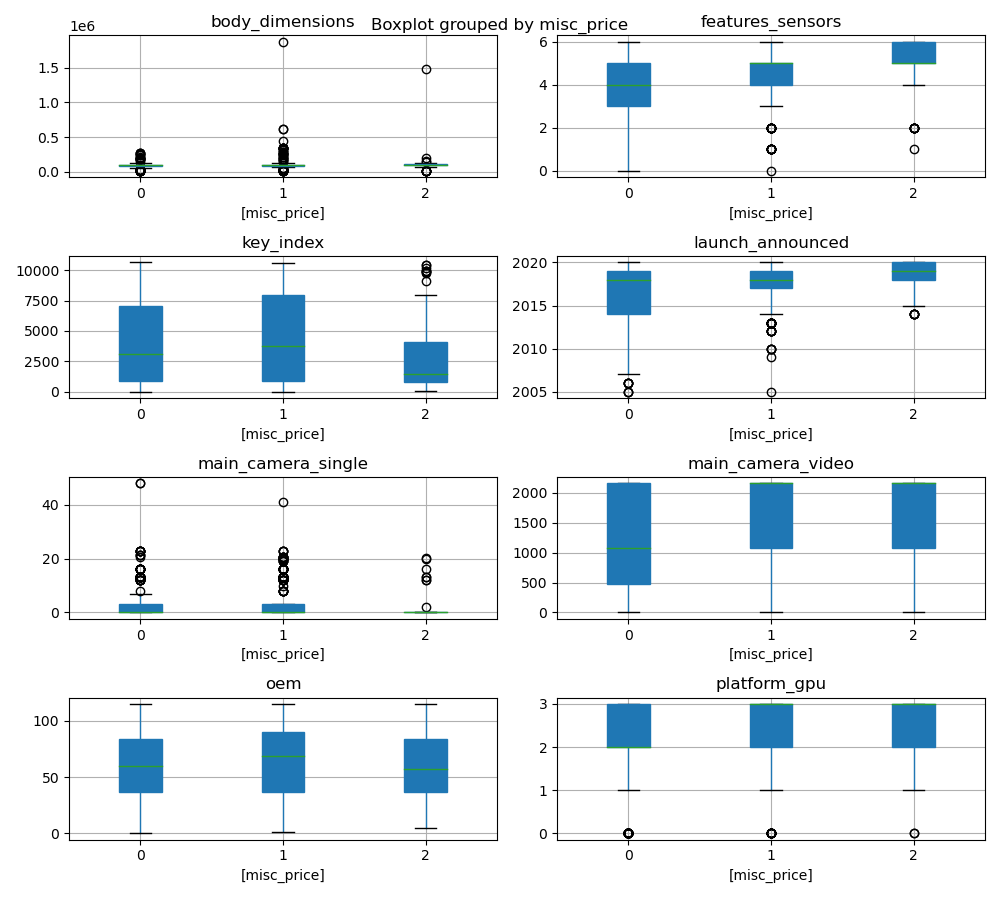


Figure : Boxplots of features grouped by price range

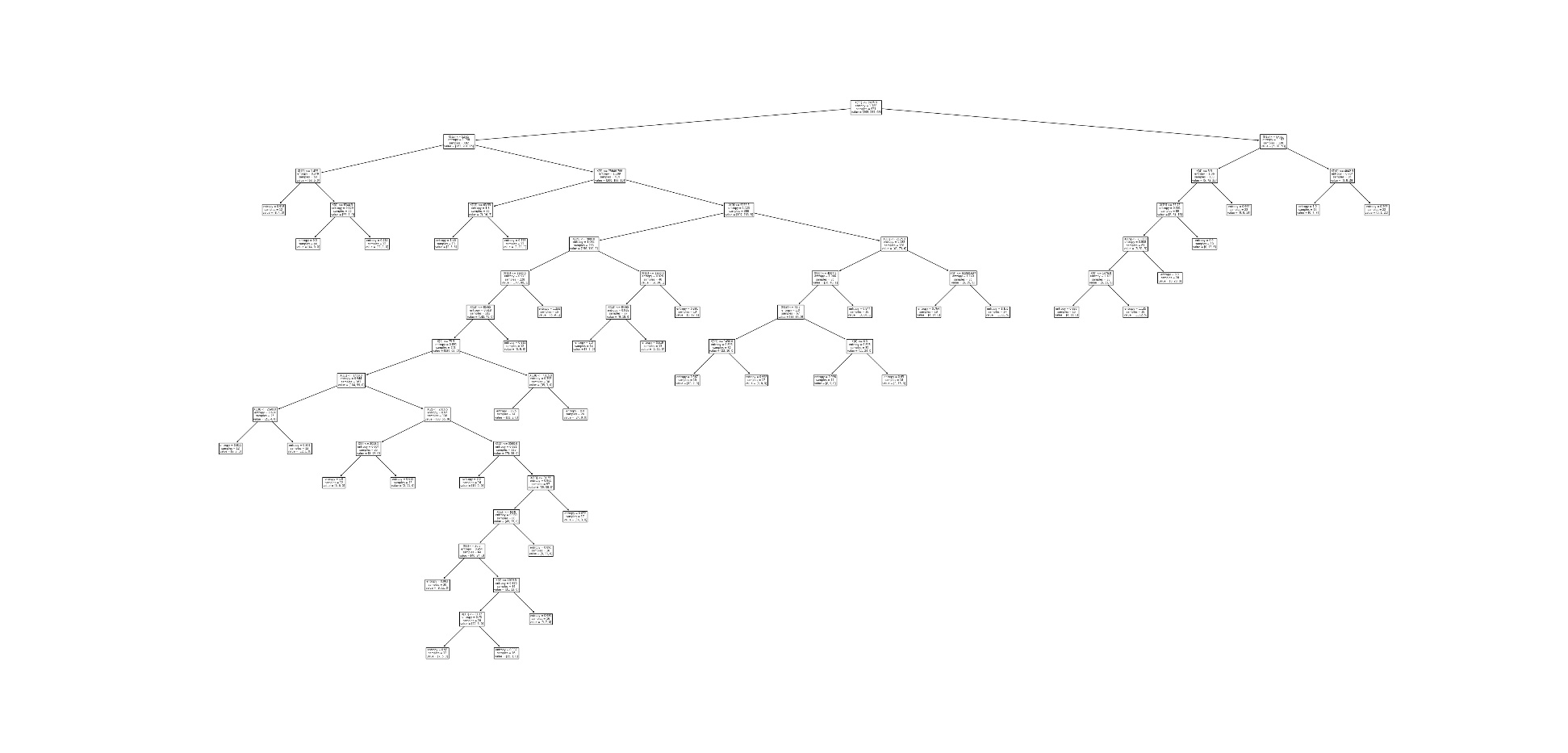
A close up of a map

Description automatically generated

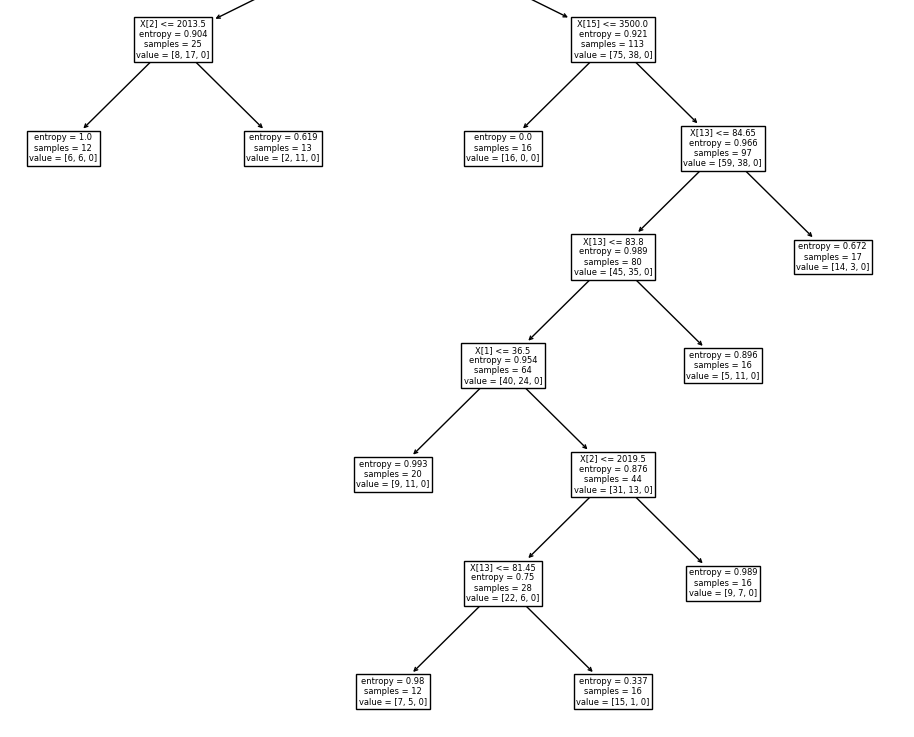
Figure : Boxplots of features grouped by price range

# Decision Tree Results

Running the default decision tree yielded an accuracy of 65%. Opting for a randomized hyperparameter search, we reached an accuracy of 68.84%. The decision tree is shown below.



A smaller branch of the tree is shown below for clarity:



# Bibliography

Asim, M. & Khan, Z., 2018. Mobile Price Class prediction using Machine Learning Techniques. *International Journal of Computer Applications,* 179(29).

Nasser, I. M. & Al-Shawwa, M., 2019. ANN for Predicting Mobile Phone Price Range. *IJAISR,* 3(2), pp. 1-6.

Saiteja, A., 2019. *Mobile Price Prediction Using Machine Learning Classification Techniques,* s.l.: s.n.