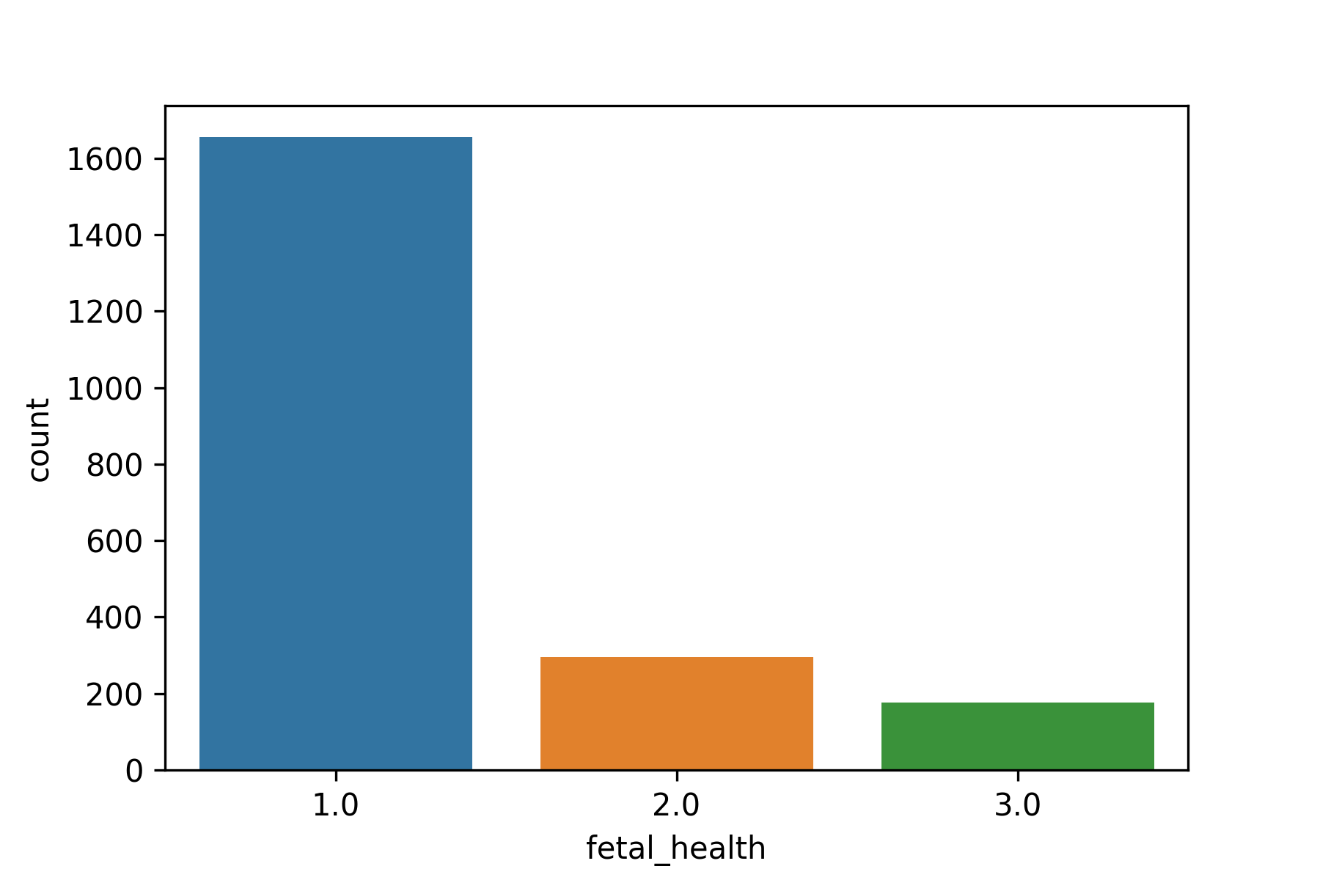
 Present a visual  distribution of the 3 classes. Is the data balanced? How do you plan to circumvent the data imbalance problem, if there is one?



We solved this by creating a new data set that oversampled from minority classes.



 Present  10 features that are most reflective to fetal health conditions (there are more than one way of selecting features and any of these are acceptable) . Present if the correlation is statistically significant (using 95% and 90% critical values).

Used Pearson correlation coefficient to select 10 features most reflective to fetal health conditions. Results as follows :

|  |  |  |
| --- | --- | --- |
| Feature | (Correlation coefficient)2 | p-value |
| abnormal\_short\_term\_variability | 0.277334 | 0 |
| prolongued\_decelerations | 0.248739 | 3.88e-297 |
| accelerations | 0.245207 | 1.24e-309 |
| histogram\_mean | 0.194471 | 3.41e-208 |
| histogram\_mode | 0.184111 | 5.95e-204 |
| histogram\_median | 0.179731 | 1.85e-192 |
| mean\_value\_of\_long\_term\_variability | 0.174783 | 7.00e-181 |
| histogram\_variance | 0.107088 | 7.35e-108 |
| percentage\_of\_time\_with\_abnormal\_long\_term\_variability | 0.075500 | 6.61e-96 |
| histogram\_tendency | 0.068981 | 4.19e-71 |

 Develop two different  models to classify CTG features into the three fetal health states (I intentionally did not name which two models. Note that this is a **multiclass** problem that can also be treated as regression, since the labels are numeric.)

We split the data into testing and training sets (test set was 30% of available data) and trained a decision tree classifier model and a logistic regression model to classify the CTG features.

** Visually present the confusion matrix (1)**

**Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted Class 1 | Predicted Class 2 | Predicted Class 3 | totals |
| Actual Class 1 | 477 | 22 | 3 | 502 |
| Actual Class 2 | 0 | 492 | 0 | 492 |
| Actual Class 3 | 0 | 0 | 496 | 496 |
| totals | 474 | 517 | 499 | 1490 |

**Logistic Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted Class 1 | Predicted Class 2 | Predicted Class 3 | totals |
| Actual Class 1 | 421 | 62 | 19 | 502 |
| Actual Class 2 | 79 | 359 | 54 | 492 |
| Actual Class 3 | 17 | 64 | 415 | 496 |
| totals | 474 | 517 | 499 | 1490 |

** With a testing set of size of 30% of all available data, calculate (1.5)**

ROC Curve AUC:

Logistic Regression : 0.91

Decision Tree : 0.99

F1 Score:

Logistic Regression : 0.79

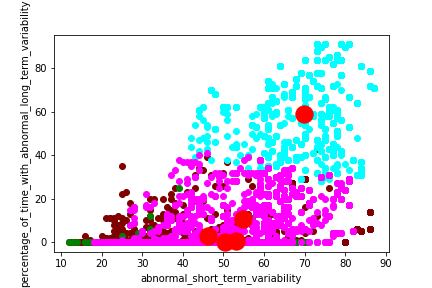
Decision Tree : 0.98

AUPRC:

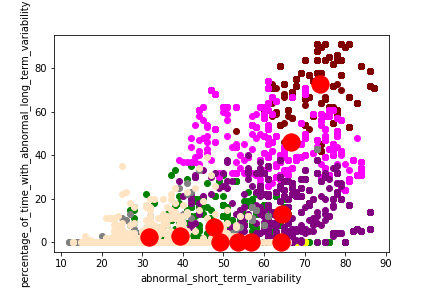
Logistic Regression : 0.86

Decision Tree : 0.97

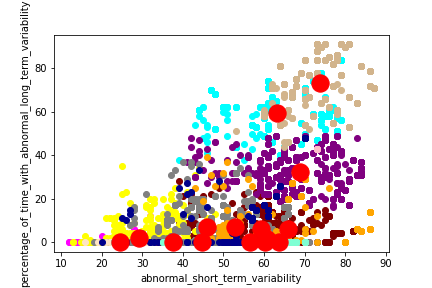
** Without considering the class label attribute, use k-means clustering to cluster the records in  different clusters and visualize them (use k to be 5, 10, 15). (2.5)**

****

k=5



k=10



k=15