SHAP Compare

Only 20 frequencies

Original dataset

The force plot is an interpretation of the predictions for a single sample which visualizes shap values as forces, each feature value is a force that increases or decreases the prediction, the prediction starts from a baseline which is the constant of the explanatory model, and each attributed value is an arrow that increases (positive) or decreases (negative) the prediction. Features that push the prediction higher (to the right) are shown in red and those that push the prediction lower are shown in blue.

Figure 1: The data in the eighth row of the predicted output is known to be 0, and the true value is also 0, so the prediction is correct. Its model predicted value is 0.06 close to 0. The figure can be interpreted as the frequency f(2378.4142) has the largest positive contribution, followed by f(408.4789), but f(865.5366) has a large negative contribution, followed by f(1029.3022), meaning that when the values of the frequencies f(2378.4142), f(408.4789), etc. are lower than their corresponding means, and the frequencies of f(865.5366), f(1029.3022), etc. are higher than the mean of the training data set, the probability that the sample has an ear with conductive condition is only 0.06.

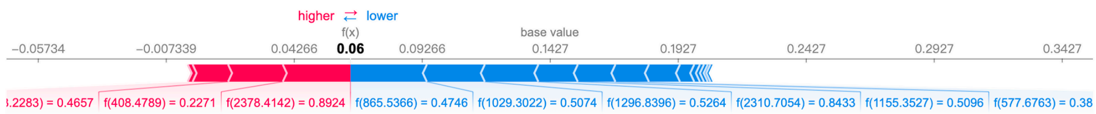
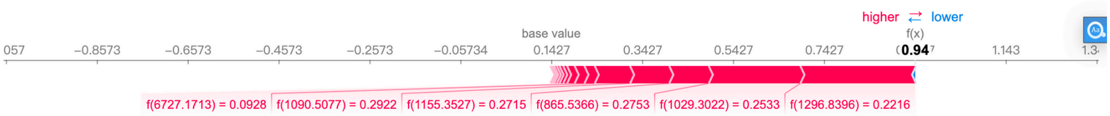
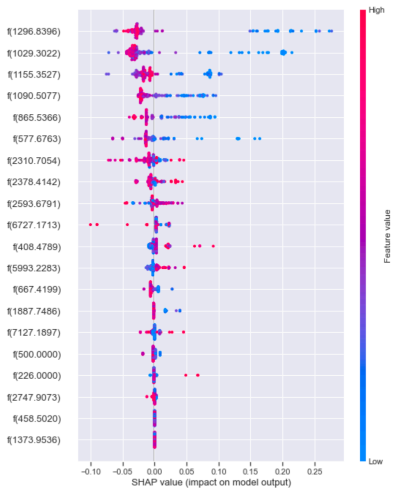


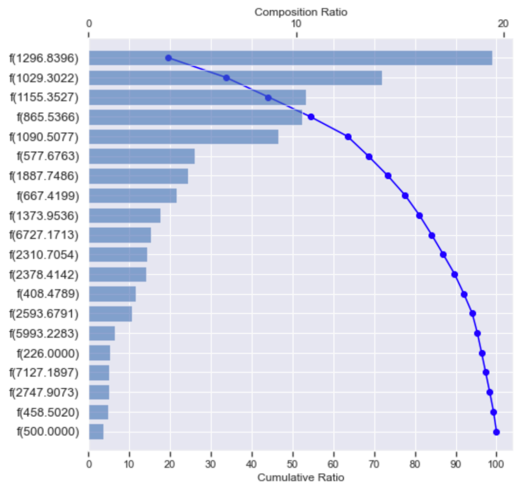
Figure 2: Predicting the seventh row of data with a known correct prediction and a result of 1, which means this ear with conductive condition. It is clear from the figure that most of the frequencies have values below their mean in the training data set, so substantially pushing the prediction to the right, resulting in a sample with an ear with the probability of the sample having an ear with conductive condition is as high as 0.94, so the prediction is biased towards 'anomaly'.



The above are all interpretations for individual instances; the next plots are for all instances.

First is the SHAP value plot, which further shows the positive and negative relationship between the predictor variables and the target variables to get the magnitude of the effect of the feature on the prediction, which is equivalent to rotating each feature of the force plot by 90 degrees and drawing it as a point plot. The effect of the size of the feature on the results is indicated by the colour (red for large values, blue for small values and purple adjacent to the mean), the wider the distribution of the area the greater its influence, thus frequencies such as f(1296.8396), f(1029.3022), f(1155.3527) have a high and positive influence, 'high ' from the red and 'negative' influences shown on the x-axis. Similarly we would say that f(2378.4142), f(2593.6791) is positively associated with diagnosing patients ear with conductive condition.





For a better understanding, let's look at the dependency diagram:
Based on the SHAP waterfall plot, we can say that f(1296.8396) is the most important frequency in the model, which almost reach 20% of the model's explainability. Also, these top 20 frequencies have provided 100% of the model's interpretation.

Oversampling dataset

The force plot is an interpretation of the predictions for a single sample which visualises shap values as forces, each feature value is a force that increases or decreases the prediction, the prediction starts from a baseline which is the constant of the explanatory model, and each attributed value is an arrow that increases (positive) or decreases (negative) the prediction. Features that push the prediction higher (to the right) are shown in red and those that push the prediction lower are shown in blue.
Figure 1: Predicting the 8th row of data, the predicted output is known to be 0, and the true value is also 0, so the prediction is correct. The predicted value of the model is 0.06 close to 0, which can be interpreted by the figure as only the frequency f(1887.7486) makes a positive contribution to the result, with f(1029.3022) making a large negative contribution, followed by f(865.5366), meaning that when the value of

f(1887.7486) is lower than their corresponding mean, and f(865.5366), f(1029.3022), etc. have frequencies above the mean of the training data set, the probability of that sample having an ear with conductive condition is 0.00.

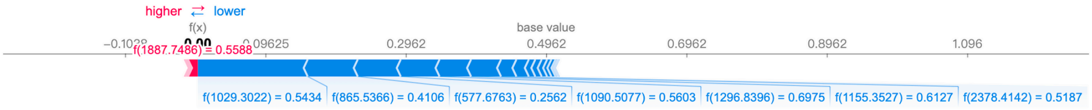
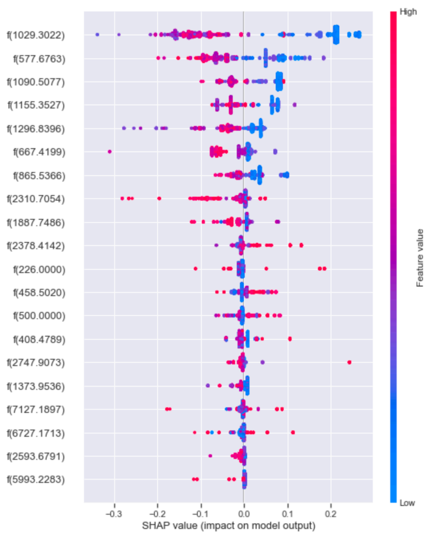
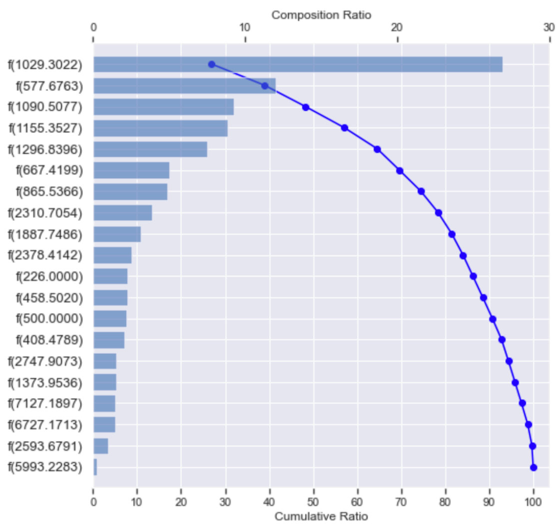


Figure 2: Predicting the 27th row of data with a known correct prediction and a result of 1, which means this ear with conductive condition. It is clear from the figure that most of the frequencies have values below their mean in the training data set, so substantially pushing the prediction to the right, resulting in a sample with an ear with The probability of the sample having an ear with conductive condition is as high as 1.0, so the prediction is biased towards 'anomaly'.

The above are all interpretations for individual instances; the next plots are for all instances.
First is the SHAP value plot, which further shows the positive and negative relationship between the predictor variables and the target variables to get the magnitude of the effect of the feature on the prediction, which is equivalent to rotating each feature of the force plot by 90 degrees and drawing it as a point plot. The effect of the size of the feature on the results is indicated by the colour (red for large values, blue for small values and purple adjacent to the mean), the wider the distribution of the area the greater its influence, thus frequencies such as f(1029.3022), f(577.6763),f(1090.5077) have a high and positive influence, 'high ' from the red and 'negative' influences shown on the x-axis. Similarly we would say that f(2378.4142), f(2593.6791) is positively associated with diagnosing patients ear with conductive condition.





For a better understanding, let's look at the dependency diagram (right), based on the SHAP waterfall plot, we can say that f(1029.3022) is the most important frequency in the model, which almost reach 30% of the model's explainability. Also, these top 20 frequencies have provided 100% of the model's interpretation.

SMOTE dataset

The force plot is an interpretation of the predictions for a single sample which visualises shap values as forces, each feature value is a force that increases or decreases the prediction, the prediction starts from a baseline which is the constant of the explanatory model and each attributed value is an arrow that increases (positive) or decreases (negative) the prediction. Features that push the prediction higher (to the right) are shown in red and those that push the prediction lower are shown in blue.

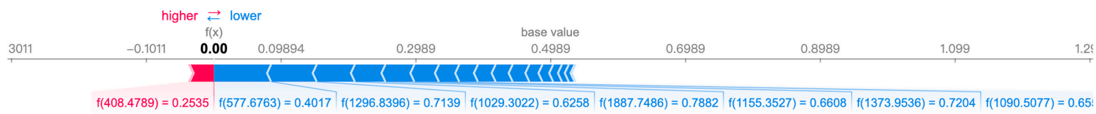
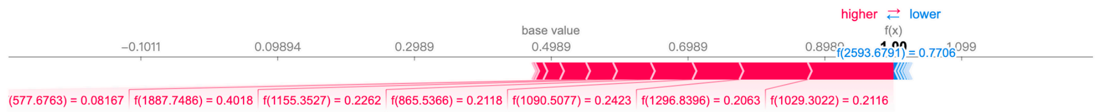
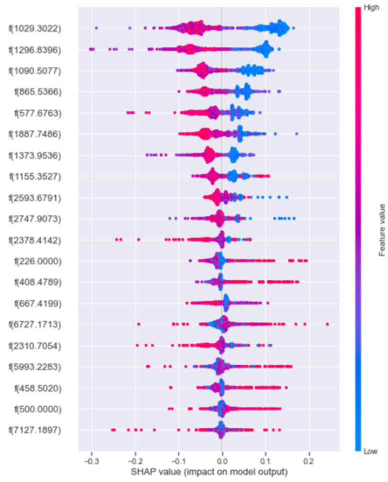


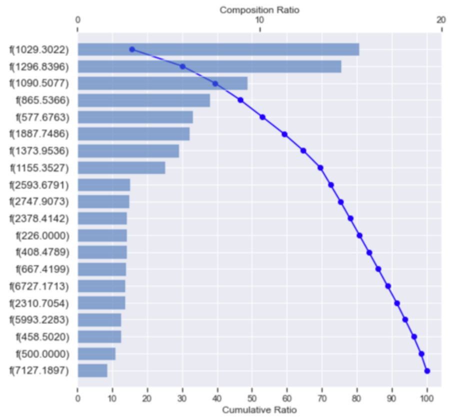
Figure 2: Predicting the second row of data with a known correct prediction and a result of 1, which means this ear with conductive condition. It is clear from the figure that most of the frequencies have values below their mean in the training data set, so substantially pushing the prediction to the right, resulting in a sample with an ear with The probability of the sample having an ear with conductive condition is as high as 1.00, so the prediction is biased towards 'anomaly'.



The above are all interpretations for individual instances; the next plots are for all instances.

First is the SHAP value plot, which further shows the positive and negative relationship between the predictor variables and the target variables to get the magnitude of the effect of the feature on the prediction, which is equivalent to rotating each feature of the force plot by 90 degrees and drawing it as a point plot. The effect of the size of the feature on the results is indicated by the colour (red for large values, blue for small values and purple adjacent to the mean), the wider the distribution of the area the greater its influence, thus frequencies such as f(1029.3022), f(1296.8396), f(1090.5077) have a high and positive influence, 'high ' from the red and 'negative' influences shown on the x-axis. Similarly we would say that f(2378.4142), f(226.0000), f(408.4789) is positively associated with diagnosing patients ear with conductive condition.





For a better understanding, let's look at the dependency diagram:
Based on the SHAP waterfall plot, we can say that f(1029.3022) and f(1296.8396) are two most important frequencies in the model, both of them almost reach 15% of the model's explainability. Also, these top 20 frequencies have provided 100% of the model's interpretation.

Adasyn datsset

The force plot is an interpretation of the predictions for a single sample which visualises shap values as forces, each feature value is a force that increases or decreases the prediction, the prediction starts from a baseline which is the constant of the explanatory model and each attributed value is an arrow that increases (positive) or decreases (negative) the prediction. Features that push the prediction higher (to the right) are shown in red and those that push the prediction lower are shown in blue.

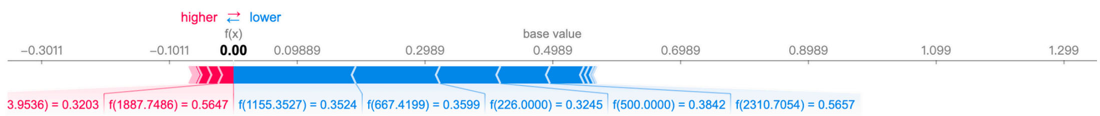
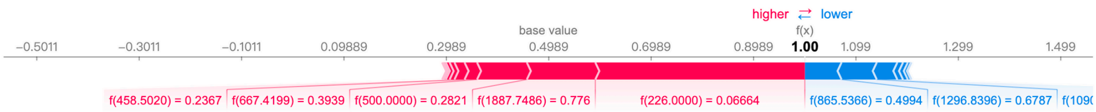
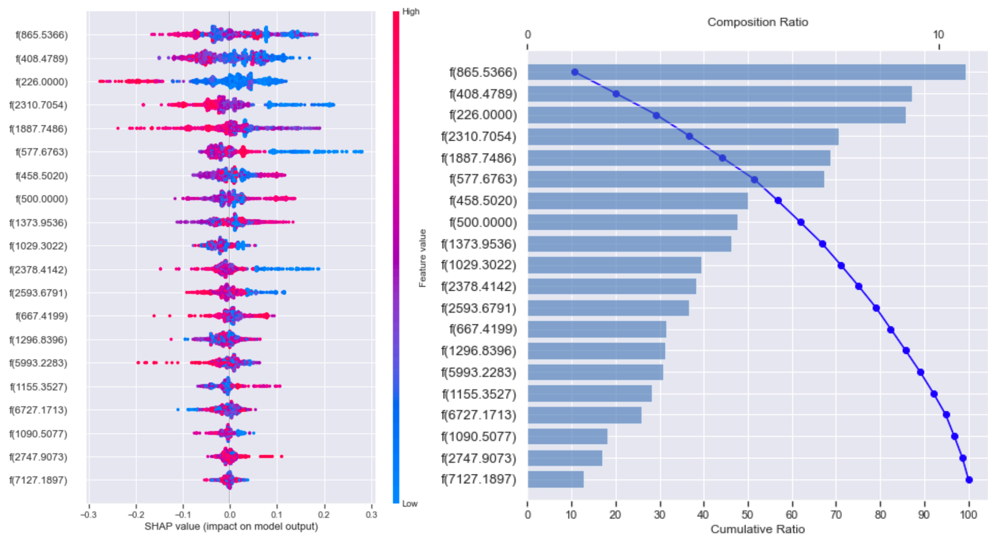


Figure 2: Predicting the second row of data with a known correct prediction and a result of 1, which means this ear with conductive condition. It is clear from the figure that most of the frequencies have values below their mean in the training data set, so substantially pushing the prediction to the right, resulting in a sample with an ear with The probability of the sample having an ear with conductive condition is as high as 1.00, so the prediction is biased towards 'anomaly'.



The above are all interpretations for individual instances; the next plots are for all instances.

First is the SHAP value plot, which further shows the positive and negative relationship between the predictor variables and the target variables to get the magnitude of the effect of the feature on the prediction, which is equivalent to rotating each feature of the force plot by 90 degrees and drawing it as a point plot. The effect of the size of the feature on the results is indicated by the colour (red for large values, blue for small values and purple adjacent to the mean), the wider the distribution of the area the greater its influence, thus frequencies such as f(865.5366),f(408.4789), f(226.0000) have a high and positive influence, 'high ' from the red and 'negative' influences shown on the x-axis. Similarly we would say that f(500.0000), f(458.5020),f(1373.9536) is positively associated with diagnosing patients ear with conductive condition.



For a better understanding, let's look at the dependency diagram (right), based on the SHAP waterfall plot, we can say that f(865.5366) is the most important frequency in the model, which is more than 10% of the model's explainability. Also, these top 20 frequencies have provided 100% of the model's interpretation.

Result Compare

Original dataset-

(max_depth=10, min_samples_leaf=2, min_samples_split=10, n_estimators=9, random_state=42)

Model		Precision	F1_score	Accuracy	Recall	Sensitivity	Specificity	roc
dt.	val	1.0	0.855	0.9310344827586207	0.8	0.6	1.0	
	test	0.75	0.8500000000000001	0.9166666666666666	0.85	0.75	0.95	0.85
	train	1.0	1.0	1.0	1.0	1.0	1.0	
rf.	val	1.0	0.9342403628117915	0.9655172413793104	0.9	0.8	1.0	
	test	1.0	0.9163763066202091	0.9583333333333334	0.875	0.75	1.0	0.88
		1.0	1.0	1.0	1.0	1.0	1.0	
rf_gs	val	1.0	0.9342403628117915	0.9655172413793104	0.9	0.8	1.0	
	test	1.0	0.9163763066202091	0.9583333333333334	0.875	0.75	1.0	0.875
	train	0.9090909090909091	0.9128965358322452	0.9567901234567902	0.8927007299270073	0.8	0.9854014598540146	
rf_random	val	1.0	0.9342403628117915	0.9655172413793104	0.9	0.8	1.0	
	test	1.0	0.9163763066202091	0.9583333333333334	0.875	0.75	1.0	0.875
	train	0.9090909090909091	0.9128965358322452	0.9567901234567902	0.8927007299270073	0.8	0.9854014598540146	

Oversampling dataset (max_depth=20, min_samples_split=10, n_estimators=8, random_state=42)

Model		Precision	Recall	Sensitivity	Specificity	Accuracy
dt.	val	1.0	1.0	1.0	1.0	
	test	1.0	1.0	1.0	1.0	1.0
rf.	val	1.0	1.0	1.0	1.0	
	test	1.0	1.0	1.0	1.0	1.0
rf_gs	val	1.0	1.0	1.0	1.0	
	test	1.0	1.0	1.0	1.0	1.0
rf_random	val	1.0	1.0	1.0	1.0	
	test	1.0	1.0	1.0	1.0	1.0

SMOTE dataset - (max_depth=20, n_estimators=15, random_state=42)

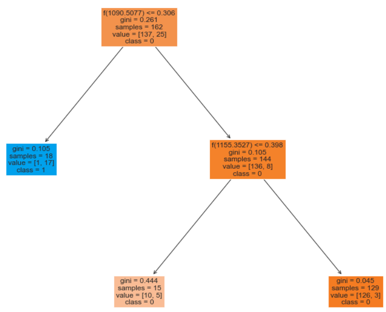
Model		Precision	Recall	Sensitivity	Specificity	Accuracy	F1_score	roc
dt.	val	0.9850980392156863	0.9850362771832761	0.9891808346213292	0.9808917197452229	0.9850980392156863	0.9890131128651831	
	test	0.9841059602649007	0.9873333333333334	0.9906666666666667	0.984	0.9873333333333333	0.9873331925910289	0.987
rf.	val	0.9921011058451816	0.9950174827502642	0.9984101748807631	0.9916247906197655	0.9951060358890701	0.9945067626829156	
	test	0.9973226238286479	0.9953333333333333	0.9933333333333333	0.9973333333333333	0.9953333333333333	0.9953333146665919	0.995
rf_gs	val	0.9969088098918083	0.9915089600969138	0.9862385321100917	0.9967793880837359	0.9913725490196078	0.9913689598873565	
	test	0.9946452476572959	0.9926666666666667	0.9906666666666667	0.9946666666666667	0.9926666666666667	0.992666637333216	0.993
	train	1.0	0.999860956618465	0.9997219132369299	1.0	0.9998615916955017	0.9998615886303365	
rf_random	val	0.9953846153846154	0.9922328591056153	0.9892966360856269	0.9951690821256038	0.9921568627450981	0.9921528030802864	
	test	0.9959623149394348	0.9913333333333334	0.9866666666666667	0.996	0.9913331445884822	0.9913331445884822	0.991
	train	0.9994418085403294	0.9976387912720144	0.9958286985539488	0.9994488839900799	0.9976470588235294	0.9976469634410439	

Adasyn datsset - (max_depth=15, max_features='sqrt', min_samples_leaf=2, min_samples_split=10, n_estimators=17, random_state=42)

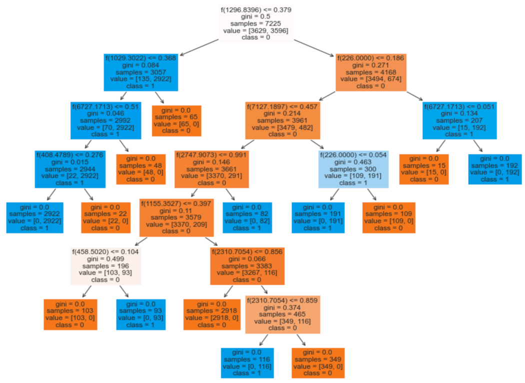
Model		Precision	F1_Score	Accuracy	Recall	Sensitivity	Specificity	Roc
dt.	val	0.9846390168970814	0.9882413292089929	0.9882445141065831	0.9881935230232444	0.9922600619195047	0.9841269841269841	
	test	0.9933422103861518	0.9933377718597425	0.9933377748167888	0.9933377718597425	0.9933422103861518	0.9933333333333333	0.993
	train	1.0	1.0	1.0	1.0	1.0	1.0	
rf.	val	0.9923076923076923	0.995296650153338	0.9952978056426333	0.9952577522236965	0.9984520123839009	0.9920634920634921	
	test	0.9933862433862434	0.9966688341807353	0.9966688874083944	0.9966666666666666	1.0	0.9933333333333333	0.997
	train	1.0	1.0	1.0	1.0	1.0	1.0	
rf_gs	val	0.9881422924901185	0.9913689598873565	0.9960815047021944	0.9960514030173473	0.9984520123839009	0.9936507936507937	
	test	0.9881422924901185	0.993337535287538	0.9933377748167888	0.9933342210386151	0.9986684420772304	0.988	0.993
	train	0.9997226844148641	0.9998616102533741	0.9998616108497094	0.999861916597625	1.0	0.9997238331952499	
rf_random	val	0.9923076923076923	0.995296650153338	0.9952978056426333	0.9952577522236965	0.9984520123839009	0.9920634920634921	
	test	0.9907651715039578	0.995336309892376	0.9953364423717521	0.9953333333333334	1.0	0.9906666666666667	0.996
	train	1.0	0.9997232199819699	0.9997232216994187	0.9997226074895977	0.9994452149791956	1.0	

Tree Diagram

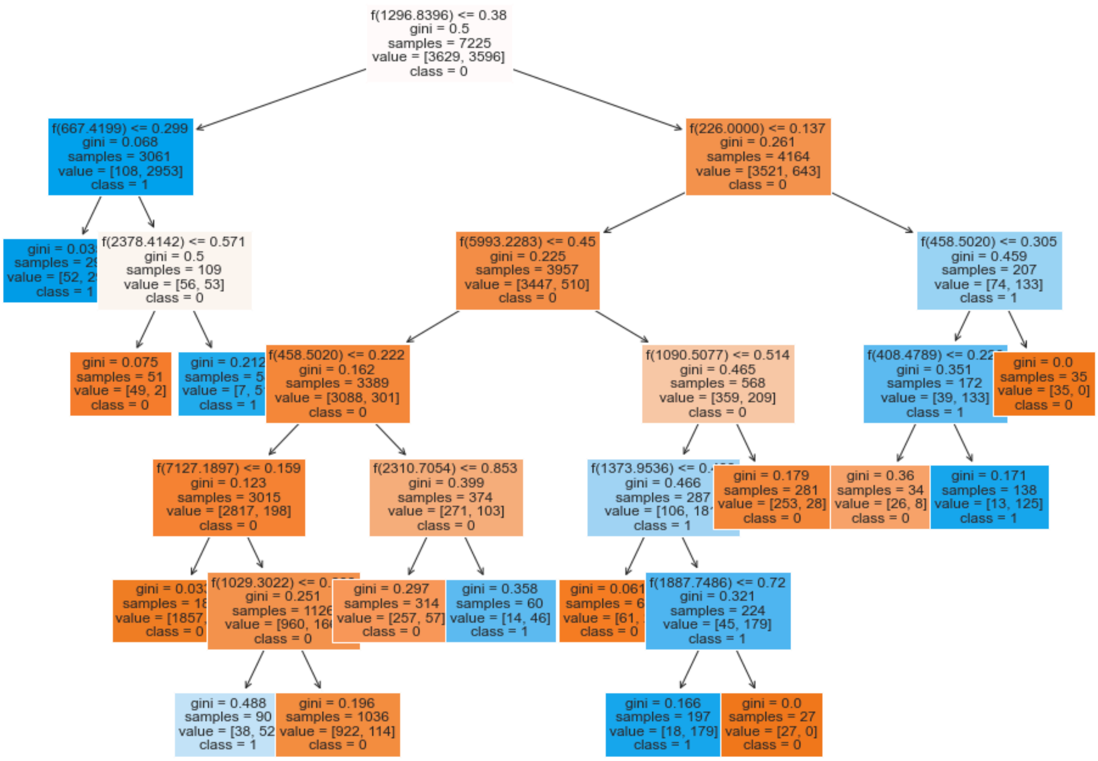
Original dataset

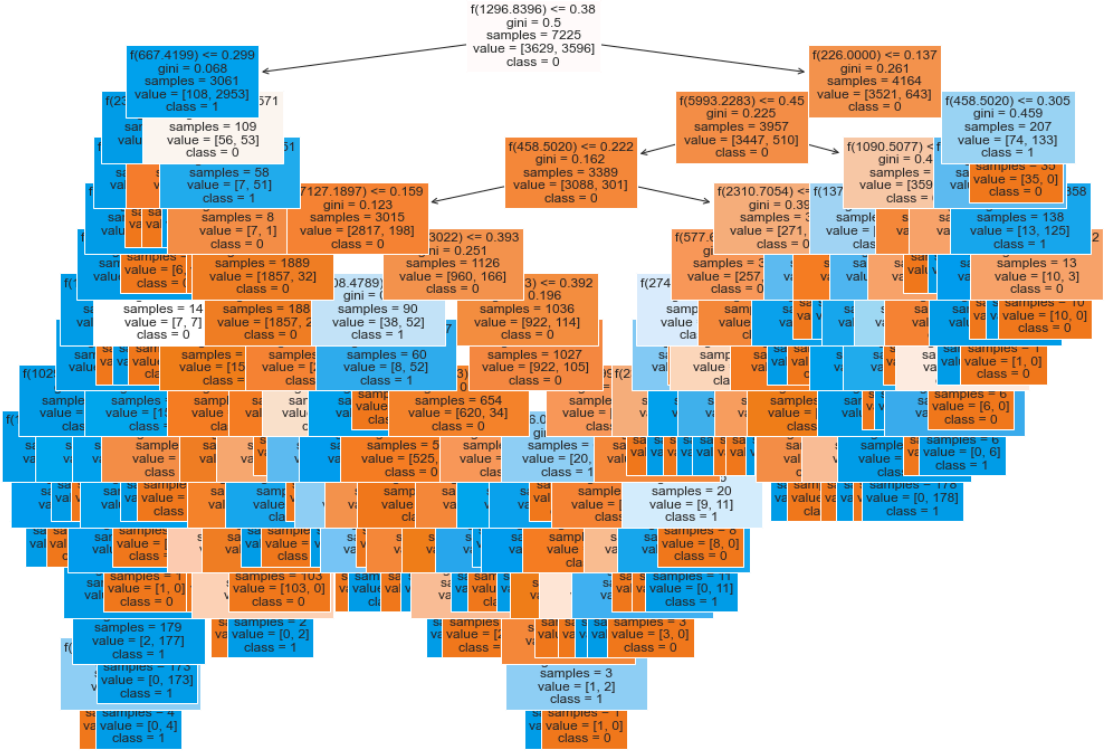


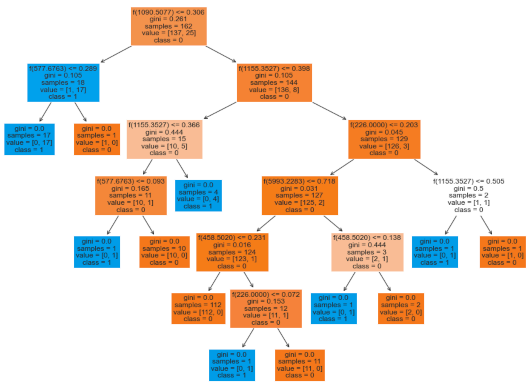
Oversampling dataset

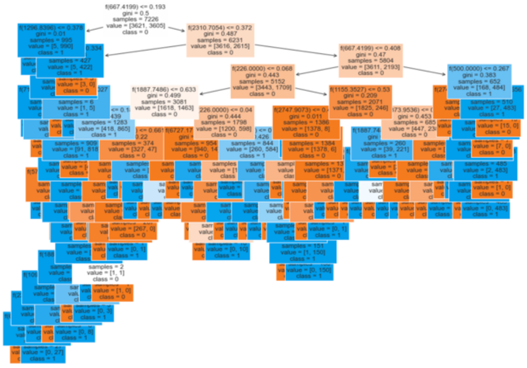


SMOTE dataset




f(1296.8396) <= 0.38
gini = 0.5
samples = 7225
value = [3629, 3596]
class = 0
f(667.4199) <= 0.299
gini = 0.068
samples = 3061
value = [108, 2953]
class = 1
f(226.0000) <= 0.137
gini = 0.261
samples = 4164
value = [3521, 643]
class = 0
f(5993.2283) <= 0.45
gini = 0.225
samples = 3957
value = [3447, 510]
class = 0
f(458.5020) <= 0.305
gini = 0.459
samples = 207
value = [74, 133]
class = 1
f(458.5020) <= 0.222
gini = 0.162
samples = 3389
value = [3088, 301]
class = 0




f(667.4199) <= 0.193
gini = 0.5
samples = 7226
value = [3621, 3605]
class = 0