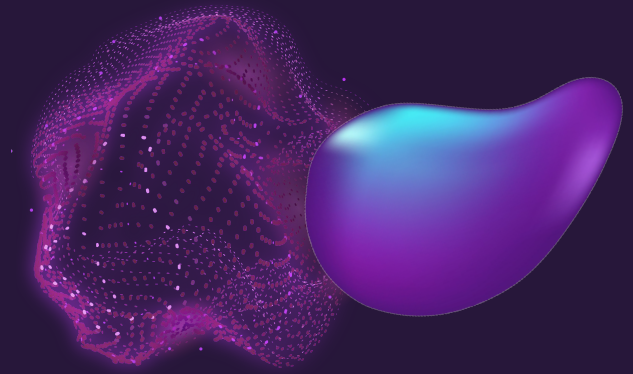


CLASSIFICATION WITH CLASS IMBALANCE PROBLEM

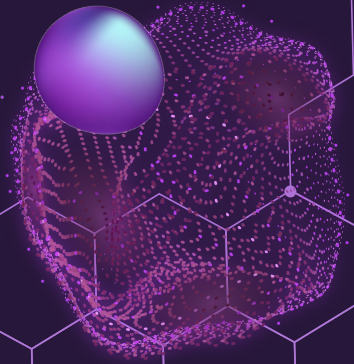
Team 58

Swetha Vipparla
Ayush Agrawal

Shubh Agarwal
Rohan Madineni

RECAP

TILL MID EVALS



CHALLENGES WITH CLASS IMBALANCE CLASSIFICATION

The classifiers fit the imbalanced data such that the accuracy of the classifier is high. The issue with this method is that, in imbalanced datasets, even if all the positive points are classified wrongly, our classifier will have a high accuracy.

Another issue with imbalanced datasets is the overlapping issue. This is the degree of separability between the classes present. When some patterns are difficult to separate, standard classifiers assign them to the negative class to increase their accuracy.

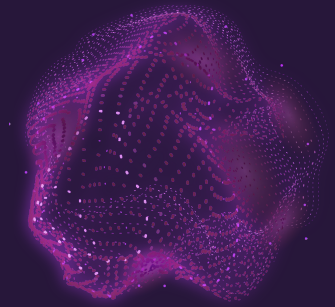
Recall is the method we check for imbalanced datasets. This is the number of samples correctly classified by the total number of correct samples.
The recall rate increases with the number of samples in the training data. Moreover, different classifiers perform differently, with some performing better than others on the same dataset. Ensemble methods are usually the best classifiers.

SAMPLING

Oversampling

This technique works on the positive class. Here, we increase the number of positive points to get a more balanced dataset. We can do this in two ways:

- Randomly select points from the minority class with replacements and add them to the training dataset
- Implement informed oversampling with methods such as SMOTE and ADASYN. These techniques basically create artificial data based on our feature space. We interpolate near points randomly. ADASYN is nearly similar but here we get more samples.



Undersampling

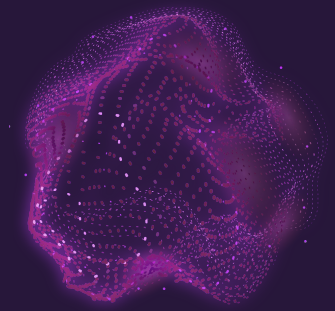
This technique works on the negative class. Here, we reduce the data points used to fit the classifier to make the classes more balanced. We can do this in two ways:

- Randomly choose the same number of data points in the negative class as the positive class
- Only consider the points in the negative class which is necessary to distinguish the classes.

FEATURE SELECTION

Correlation Matrix

Here, we see how the attributes are correlated with each other. The values range from -1 to 1 with 1 meaning they are directly proportional to each other. We try to remove all those values which have a high correlation.



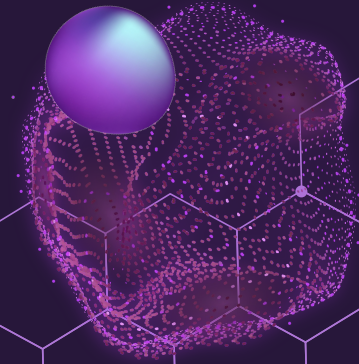
Chi-Square Test

This test is used to assess how well a predictive categorization model performs at making predictions. The outputs are numbers for each of the attributes.

We usually take the top 10 or 15 features and remove all other features. In the end, we can see that the reduction of these features were useful in increasing the recall rates.

ALGORITHM-LEVEL APPROACH FOR HANDLING CLASS IMBALANCE PROBLEM

IMPROVED ALGORITHMS

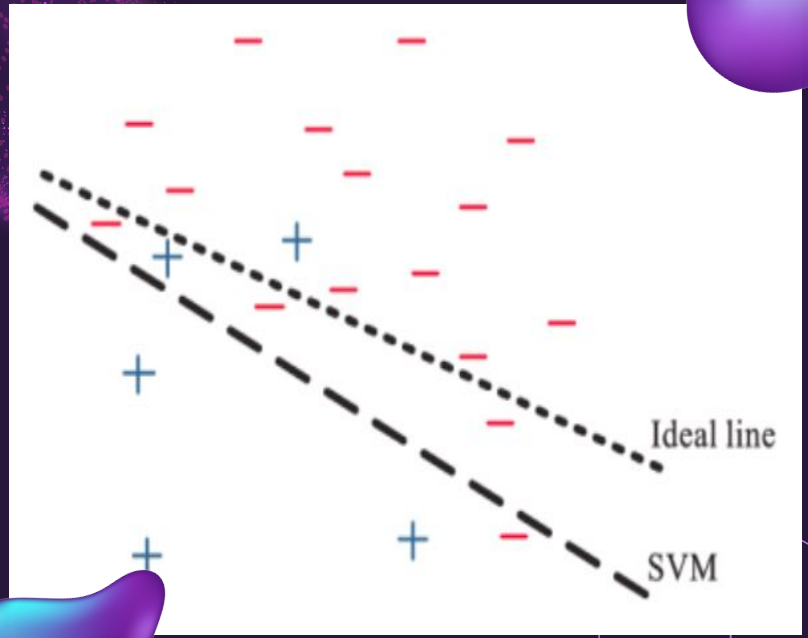


ZSVM

01	INTUITION	Idea behind the algorithm
02	APPROACH	Approach to overcoming the classic problem
03	DETERMINING Z	Use of g-mean to calculate z
04	RESULTS	Results of zsvm on our data

INTUITION

SVM is a widely used binary classification due to its accuracy and performance. It is usually able to find a discriminatory hyperplane that has a good margin for separation. But, for imbalanced data, the decision boundary is skewed towards the positive class, decreasing the recall. ZSVM concentrates on post adjustment of the decision boundary.



APPROACH

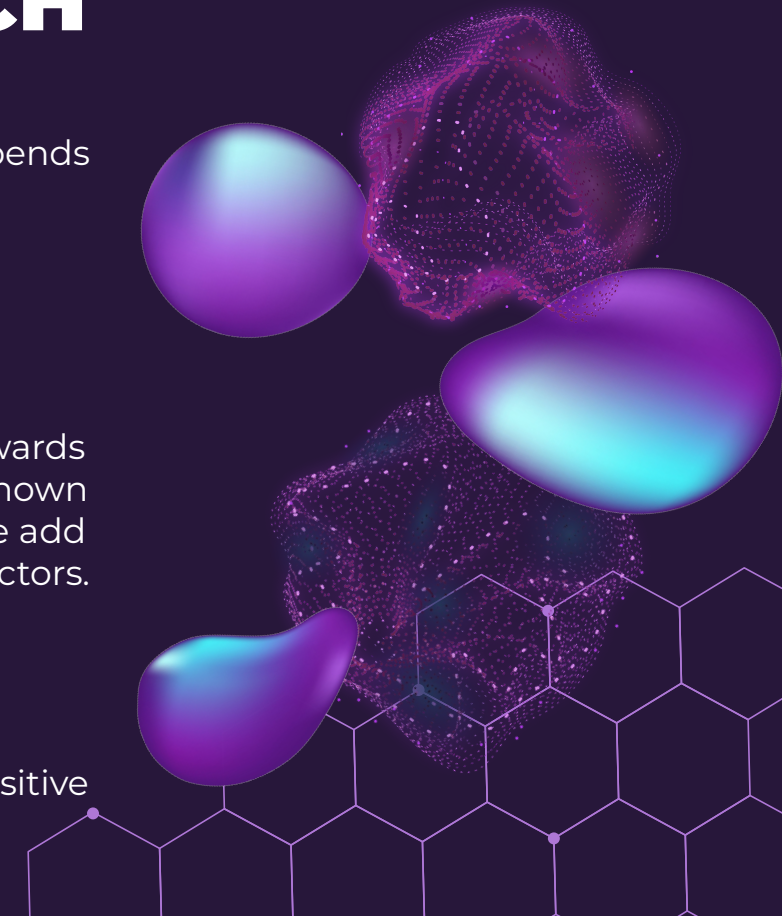
As the equation shows, the classification depends on the following factors:

- The Lagrange multiplication constant
- Kernel function
- Number of positive and negative support vectors

For imbalanced training data, these factors are biased towards the negative class, causing a tendency to classify unknown instances as negative. Thus, to decrease this tendency, we add a bias, z, with positive class support vectors.

$$f(\mathbf{x}, z) = z \sum_{\mathbf{x_p} \varepsilon SV; y_p > 0} \alpha_p y_p K(\mathbf{x}, \mathbf{x_p}) + \sum_{\mathbf{x_n} \varepsilon SV; y_n < 0} \alpha_n y_n K(\mathbf{x}, \mathbf{x_n}) + b$$

Thus, by adding the weight, we shift the leaned positive to another position further away, reducing skew.



DETERMINE Z

G-MEAN

We use G-mean to calculate the measure.

$$gmean = \sqrt{acc_+ * acc_-}$$

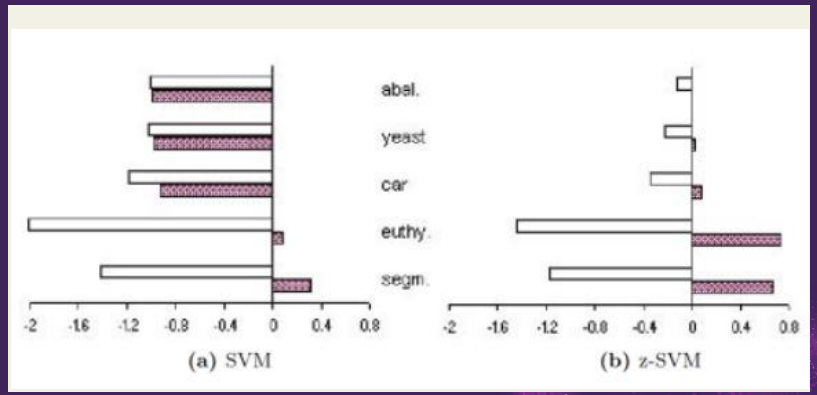
Here, gmean depends on the accuracy of both classifications of the positive as well as the negative class. So, if any of them is not optimum, it reduces the value drastically.

VALUE OF Z

At z = 0, all the points are classified as negative
while for z = m, all the points are classified as positive. Basically, we will reach a z, z > 0, and z < M such that it will be the maximum value of gmean
.

$$max_z J(z) = \sqrt{\frac{\sum_{\mathbf{x_u} \epsilon X; y_u > 0} I(y_u f(\mathbf{x_u}, z))}{P} . \frac{\sum_{\mathbf{x_v} \epsilon X; y_v < 0} I(y_v f(\mathbf{x_v}, z))}{N}}$$

RESULTS



We can see from the figure that z-SVM performs better on all imbalanced datasets. The first three datasets are extremely skewed resulting in both the distances for the hyperplane from the positive class to the negative class.

But with z-SVM, we can see that these imbalances are successfully overcome by shifting the decision boundary away from positive class.

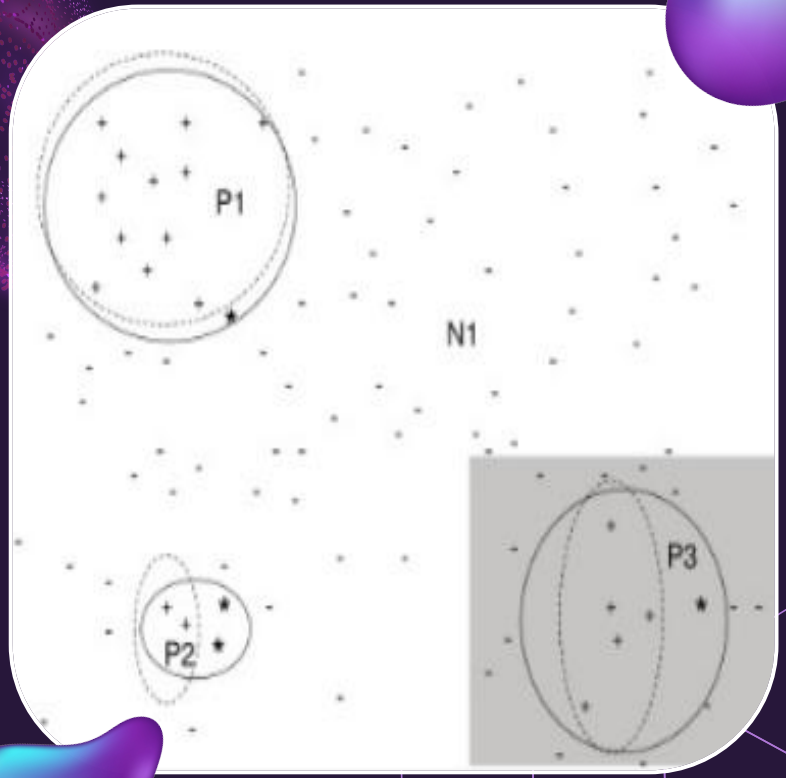
KNN EXEMPLAR GENERALIZATION

01	ISSUES WITH KNN	Issues related to class imbalance
02	DIFFERENCE	Difference between 1NN and 1ENN
03	PPI	Definition of positive pivot instances
04	ALGORITHM	kENN algorithm
05	PERFORMANCE	Performance based on other classifiers

ISSUES WITH KNN

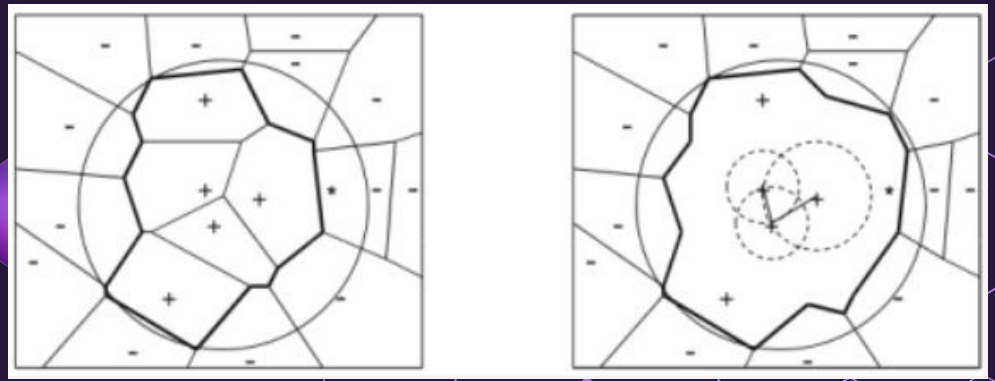
Here P1, P2, and P3 are the sub-concepts for the positive class.

Here, we can see the ideal decision boundary in the solid line, and the dashed decision boundary using in KNN. We can clearly see that there is a difference and the * points, which are the test samples are incorrectly classified.



DIFFERENCE

Here, we use 1NN. We can clearly see the difference between 1NN and 1ENN. We can see that 1ENN has expanded its decision boundary to include more points.



POSITIVE PIVOT INSTANCE

PPI

We define PPIs to increase the decision boundary.

r

r is defined as the distance between the point in question to the nearest positive neighbor e.

FP

We define False Positive Rate, FP, as the ratio of the number of negative points to all points within distance r from the point in question.

RELATION

PPI is a point whose FP rate is below a threshold. i.e. FP rate <= delta.

ALGORITHM

1	Go through each of the positive points.
2	Find the nearest positive neighbour.
3	Calculate the FP ratio.
4	If lesser than the threshold, add to the PPI list, otherwise go to step 2 for the next point.

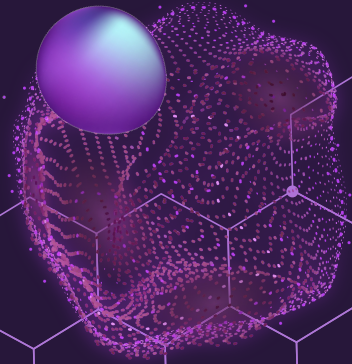
PERFORMANCE

3ENN significantly outperforms other classifiers even those built on imbalanced datasets such as the C4.5Smt+ and C4.5Meta.

Dataset	3ENN	Naive	3NN	3NNSmt+	3NNMeta	C4.5	C4.5Smt+	C4.5Meta
Oil	**0.811**	0.788	0.796	0.797	0.772	0.685	0.771	0.764
Hypo-thyroid	0.846	0.831	0.849	0.901	0.846	0.924	**0.948**	0.937
PC1	**0.806**	0.786	0.756	0.755	0.796	0.789	0.728	0.76
Glass	0.749	0.623	0.645	0.707	0.659	0.696	0.69	**0.754**
Satimage	**0.925**	0.839	0.918	0.902	**0.928**	0.767	0.796	0.765
CM1	**0.681**	0.606	0.637	0.666	0.625	0.607	0.666	0.668
New-thyroid	**0.99**	0.945	0.939	0.972	0.962	0.927	0.935	0.931
KC1	**0.794**	0.732	0.759	0.756	0.779	0.64	0.709	0.695
SPECT_F	**0.767**	0.728	0.72	0.725	0.735	0.626	0.724	0.643
Hepatitis	**0.783**	0.71	0.758	0.772	0.744	0.753	0.713	0.745
Vehicle	0.952	0.945	**0.969**	0.942	0.956	0.921	0.926	0.929
German	**0.714**	0.677	0.69	0.686	0.705	0.608	0.649	0.606
Average	**0.818**	0.768	0.786	0.798	0.792	0.745	0.771	0.766

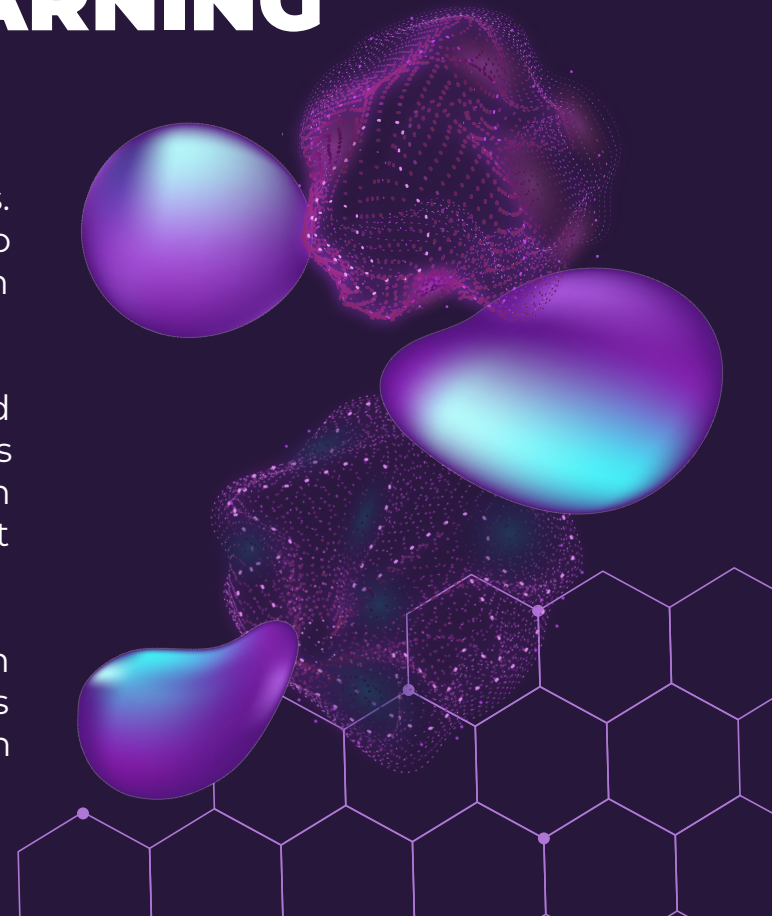
ALGORITHM-LEVEL APPROACH FOR HANDLING CLASS IMBALANCE PROBLEM

ONE CLASS LEARNING



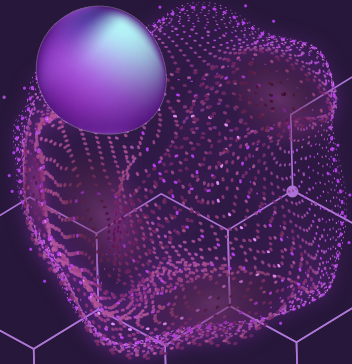
ONE CLASS LEARNING

- This algorithm attempts to focus on the issue of outliers. Using unsupervised learning, it makes an effort to simulate "normal" instances in order to categorise fresh examples as either normal or abnormal (e.g. outliers).
- Binary classification tasks with a significantly skewed class distribution can be handled using one-class classification methods. These methods can be tested on a holdout test dataset after being fitted to the input instances from the majority class in the training dataset.
- When few positive samples are provided, the algorithm can train on the negative class as the normal class. As soon as it detects an outlier, it will classify that as an anomaly. This anomaly is our positive class.



ALGORITHM-LEVEL APPROACH FOR HANDLING CLASS IMBALANCE PROBLEM

COST-SENSITIVE LEARNING



COST-SENSITIVE LEARNING

01 **INTUITION**

02 **APPROACH**

03 **RESULTS**

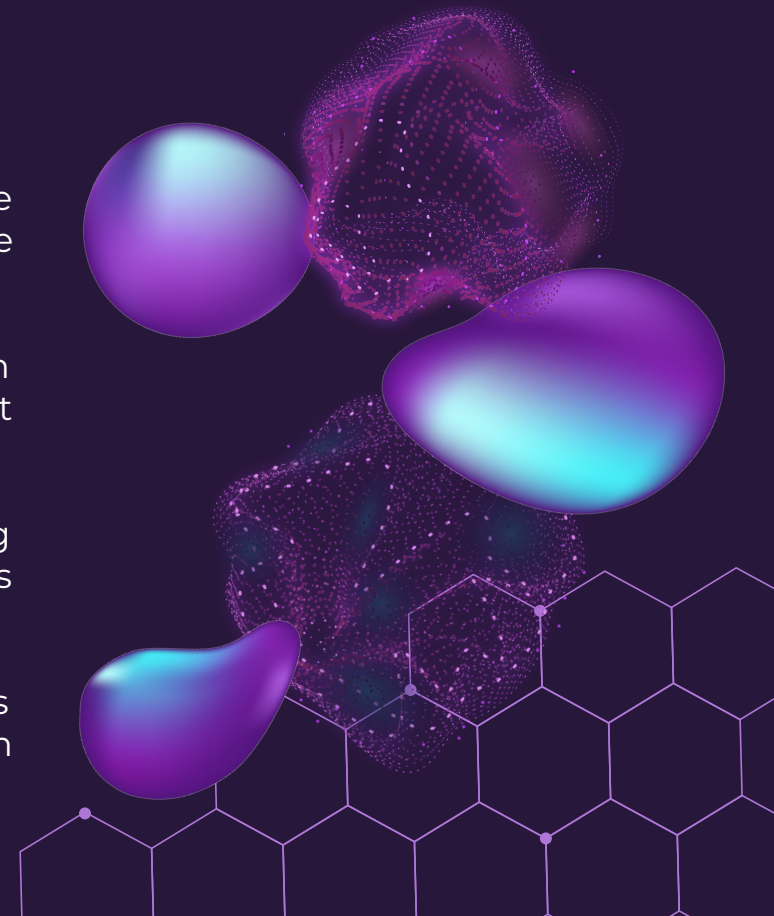
Explanation of basic need and uses of the algorithm

Approach of the algorithm

Comparison of results with and without cost sensitive approach

INTUITION

- Many classic learning algorithms view the misclassification cost as equal which is not true in the case of imbalanced data.
- Penalty-sensitive learning algorithms are created with the premise that when a classifier makes a mistake, it incurs a high cost.
- SVMs are efficient as they find the hyperplane separating the 2 classes effectively, but they struggle when the class distribution has a significant skew.
- A variant of SVM called cost-sensitive SVM incorporates this fact and enhances the performance of the algorithm on highly skewed datasets.



APPROACH

Particularly, each sample in the training dataset contains a unique penalty term (C value) that is employed in the SVM model's margin computation.

When the SVM model is fitted, specifically when the decision boundary is determined, the C parameter is employed as a penalty.

The margin can be made softer for the minority class by using a larger weighting, while the margin can be made harder and misclassified examples are avoided by using a smaller weighting for the majority class.

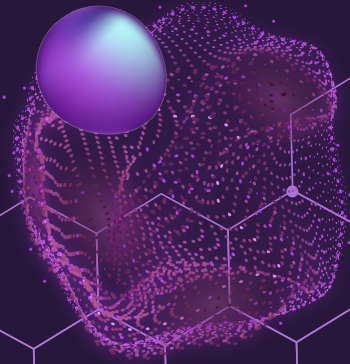
RESULTS

Without Cost-Sensitive SVM	
Measure	**Value**
ROC AUC	0.829
Recall	0.657
Precision	1.000
Specificity	1.000
G-Mean	0.811

With Cost-Sensitive SVM	
Measure	**Value**
ROC AUC	0.912
Recall	0.829
Precision	0.674
Specificity	0.996
G-Mean	0.908

ALGORITHM-LEVEL APPROACH FOR HANDLING CLASS IMBALANCE PROBLEM

ENSEMBLE METHODS



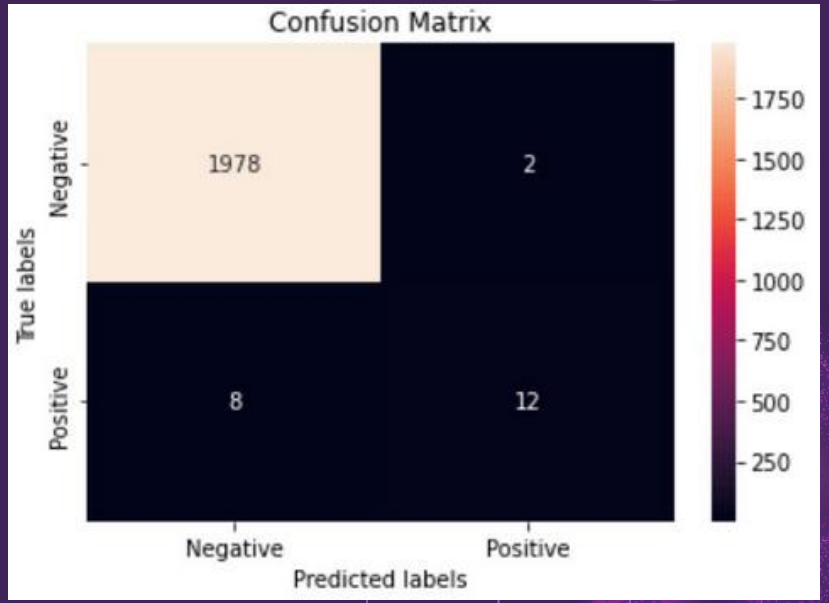
AdaBoost

Adaboost or Adaptive boosting is a boosting technique that uses a variation in the bootstrap sampling algorithm.

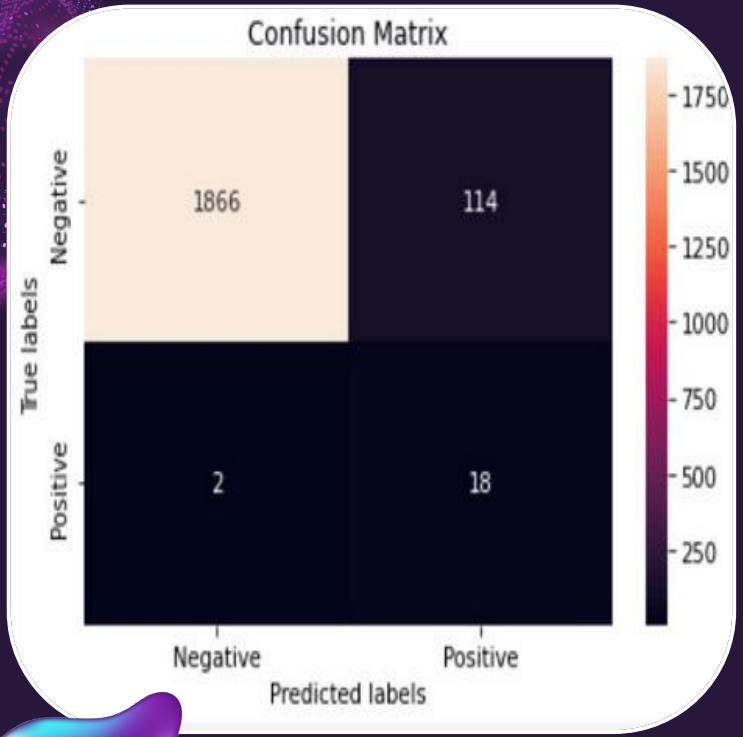
It's a variation of the bootstrap sampling algorithm where the weights assigned to samples are updated such that higher weights are placed on samples misclassified by a trained classifier model.

These magnified weights correspond to greater importance being placed on them in the subsequently trained models(such as a higher probability of being sampled).

Thus, each subsequently trained classifier is tuned to the errors of the prior ones and the result is obtained through an aggregate.



SmoteBoost



SmoteBoost utilizes the SMOTE algorithm (SyntheticOverSampling). SMOTE algorithm oversamples data by calculating synthetic instances of the data we wish to oversample and adding it to the dataset.

SMOTEBoost injects the SMOTE algorithm at each iteration of the boosting step to oversample the minority class.

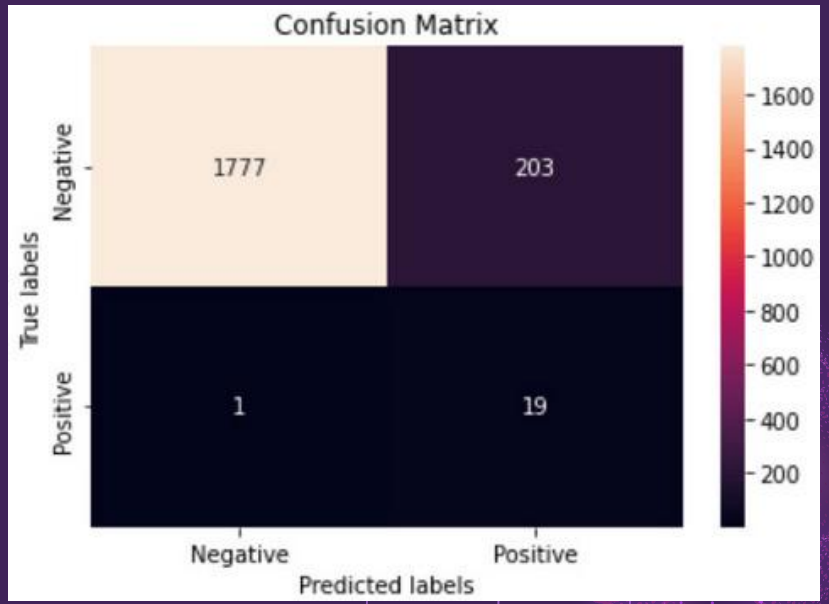
RusBoost

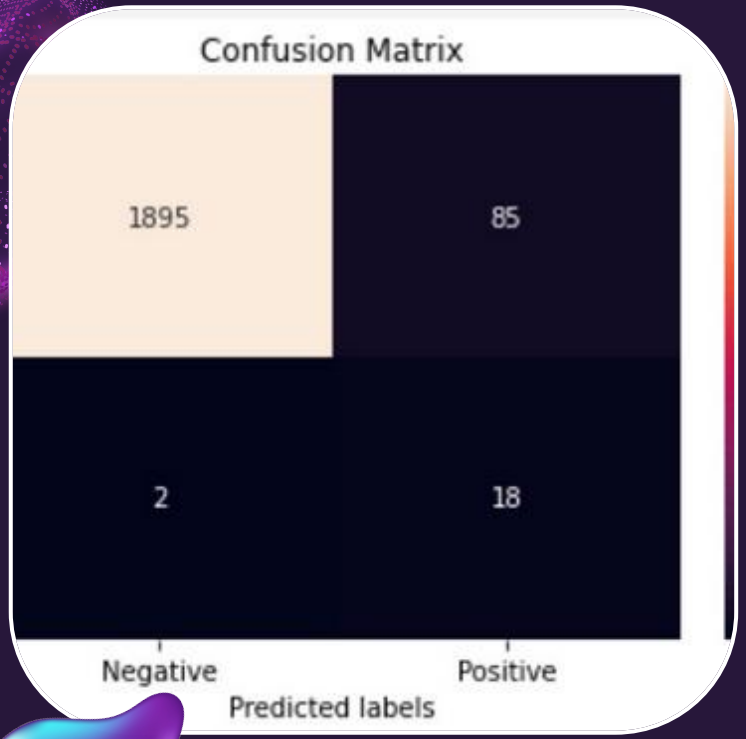
RUSBoost is a boosting-based ensemble technique similar to SMOTEBoost.

RUSBoost utilizes the RUS algorithm (Random UnderSampling). RUS undersamples data by randomly removing instances of the data we wish to undersample.

RUSBoost performs Random UnderSampling on the majority class at each boosting step.

It is faster than oversampling techniques(SMOTEBoost) and gives the highest recall of the minority class.





Under Bagging

Underbagging is a bagging-based ensemble method. It involves a combination of undersampling and bagging techniques.

It creates bags of balanced datasets during the bagging process by including all instances of the minority class dataset and randomly sampling the required number of datapoints from the majority class.

Classifier models are then trained in parallel on each of these bags and the results are obtained by an aggregation.

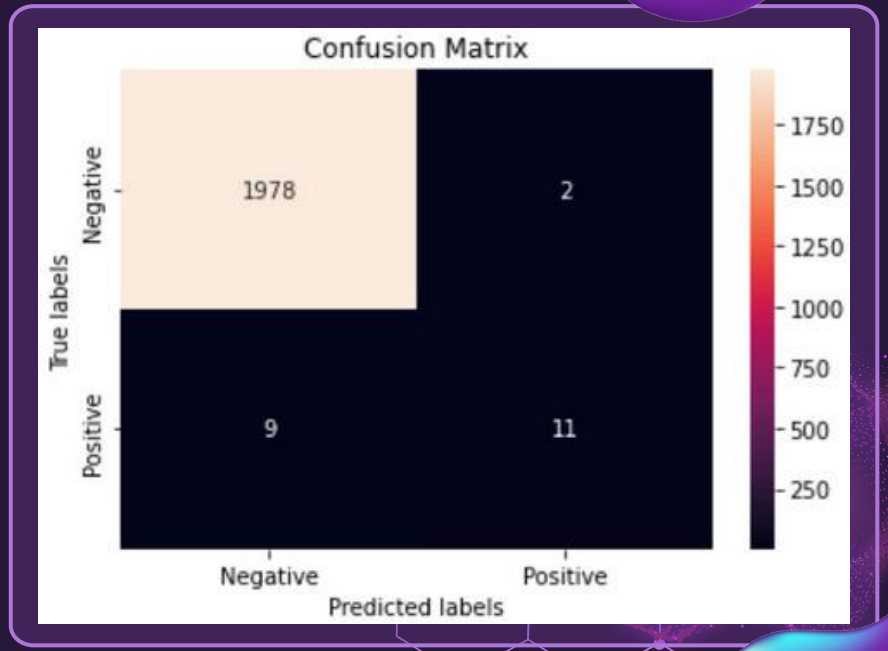
Over Bagging

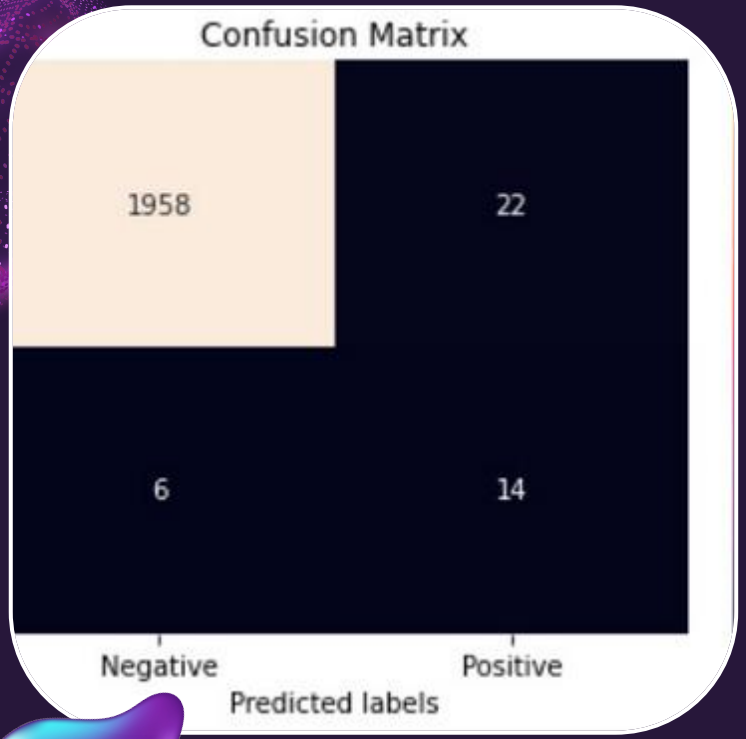
Overbagging is a bagging-based ensemble method.

It involves a combination of oversampling and bagging techniques.

Overbagging increases or adds minority class instances(oversamples) in each bag created during the bagging process.

Classifier models are then trained in parallel on each of these bags and the results are obtained by an aggregation.





SMOTE Bagging

SMOTEBagging is a bagging-based ensemble method. It involves a combination of oversampling and bagging techniques.

It involves a combination of the SMOTE and bagging algorithms. The minority class of each bag of data created is oversampled using SMOTE sampling algorithm.

Classifier models are then trained in parallel on each of these bags and the results are obtained by an aggregation.

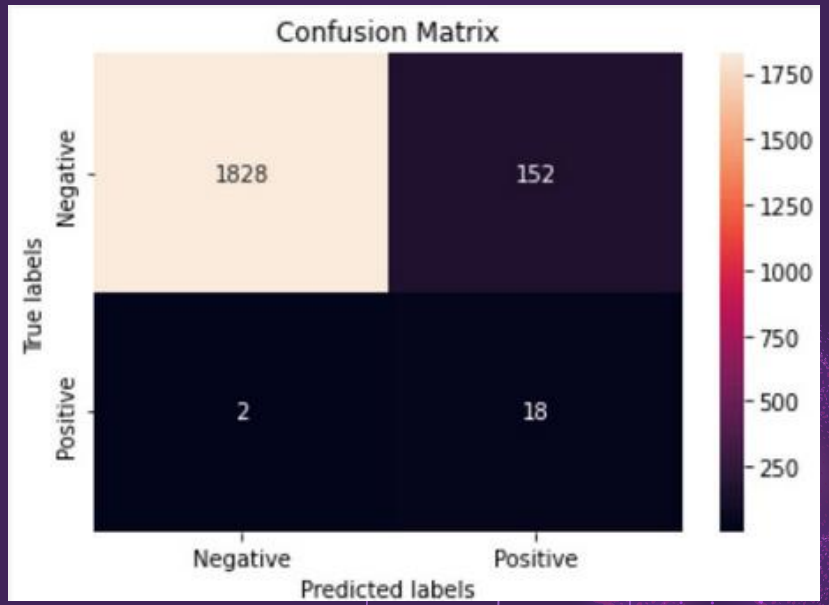
Easy Ensemble

EasyEnsemble is an undersampling algorithm that works in combination with AdaBoost.

It samples several subsets from the majority class data with each subset being equal in size to the minority class dataset.

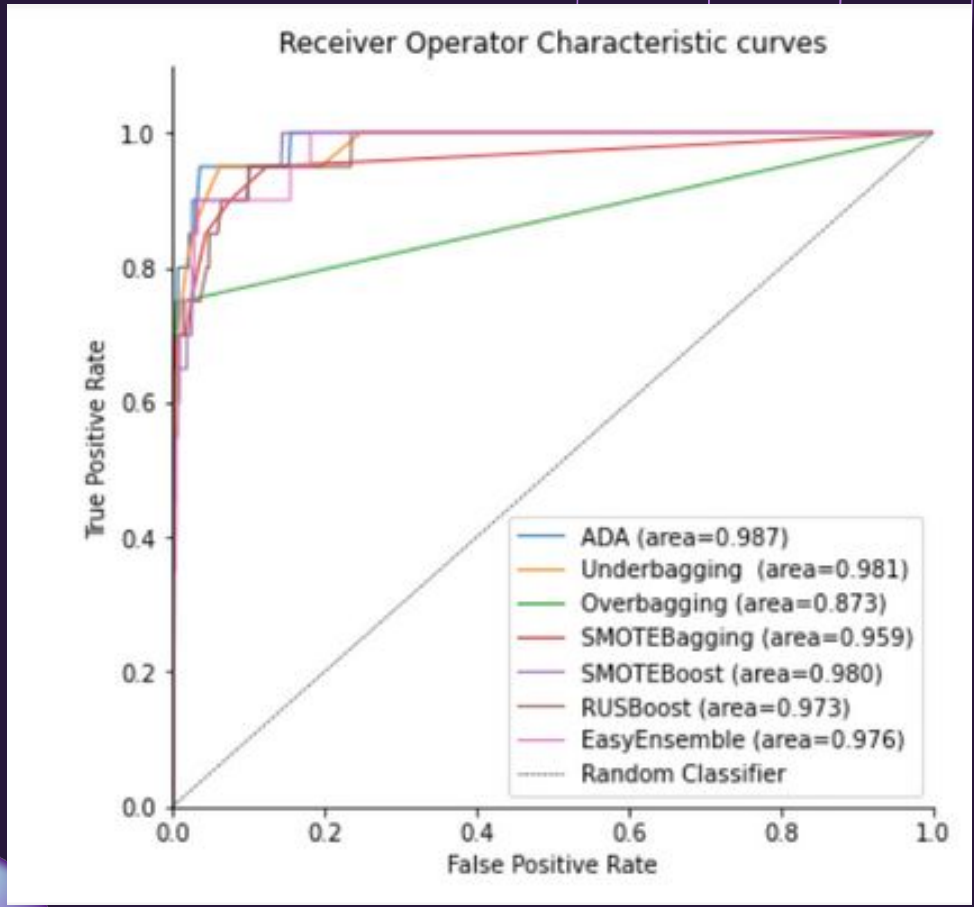
The weak classifier models are trained on a union of each of these subsets with the entire minority class dataset.

This is done in unison with AdaBoost to then obtain the final result through an aggregation.



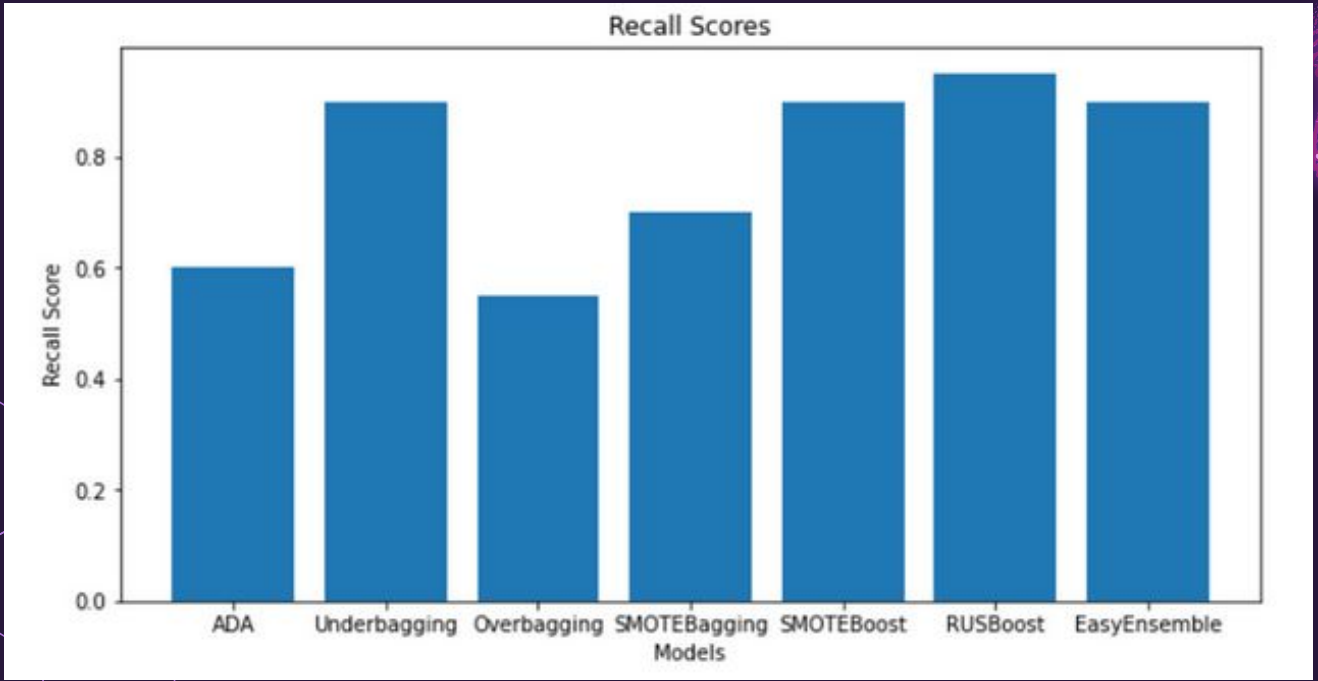
ROC-CURVE

ENSEMBLE CLASSIFIERS



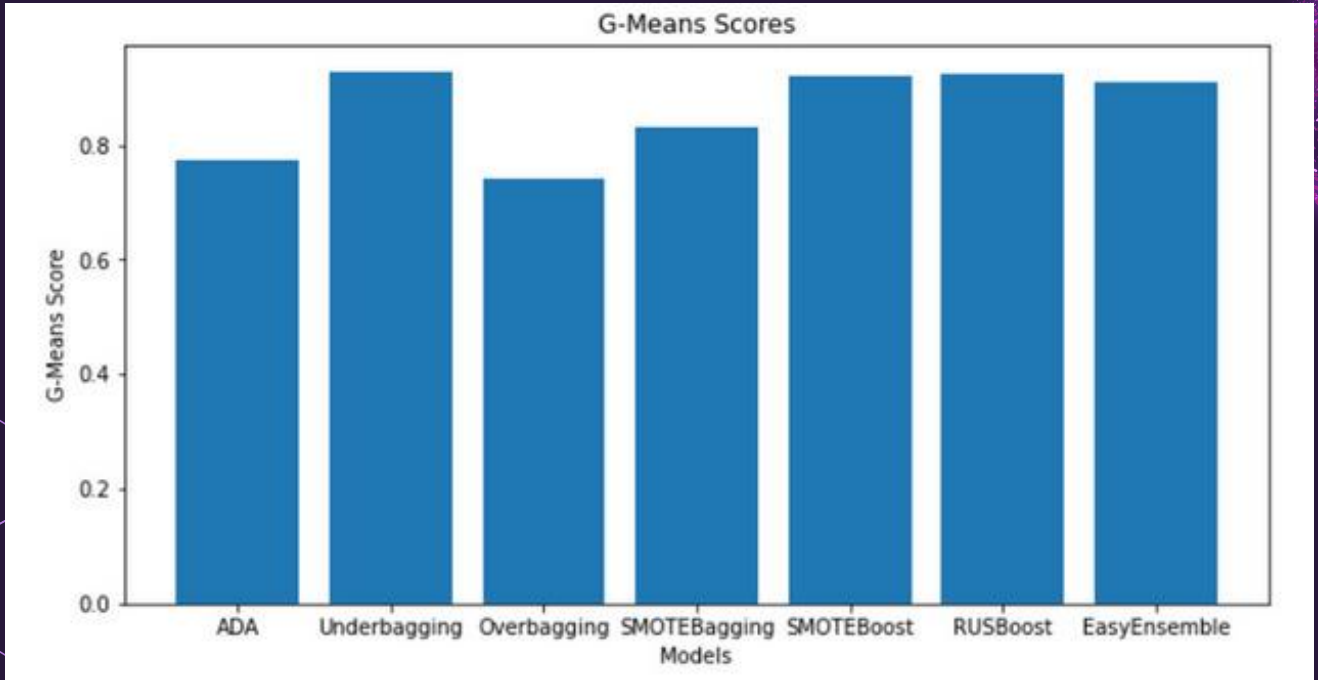
RECALL SCORES

ENSEMBLE CLASSIFIERS

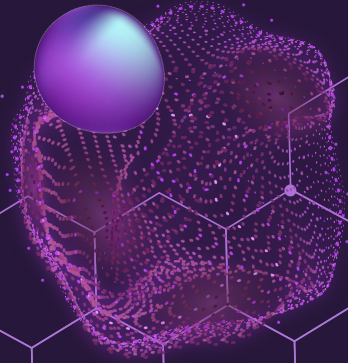


G-MEAN SCORES

ENSEMBLE CLASSIFIERS



PERFORMANCE MEASURES



PERFORMANCE MEASURES

As mentioned before, accuracy is pretty misleading for imbalanced classes. As, we have a lot of data points for one class, just by classifying all points as the negative class, we can get good accuracy.

Moreover, in most imbalanced class situations, we feel the positive class to have higher importance, i.e., we want to rightly classify them.

SPECIFICITY & SENSITIVITY

Sensitivity/True Positive Rate/Recall

It is the measure that tells us how well it classifies the positive class.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity/True Negative Rate

It is the measure that tells us how well it classifies the negative class.

$$Specificity = \frac{TN}{TN + FP}$$

These values range between 0 and 1 from the worst to the best performance.

G-Mean & Precision

G-Mean

This metric provides us with the ability to measure the balance between the positive and the negative class accuracy. This measure takes both sensitivity and specificity together. If the measure is low, then it means that the classifier is highly biased toward one class.

$$G - mean = \sqrt{Sensitivity \times Specificity}$$

Precision

It is the measure which tells that from all the predicted positives, how many were true positives.

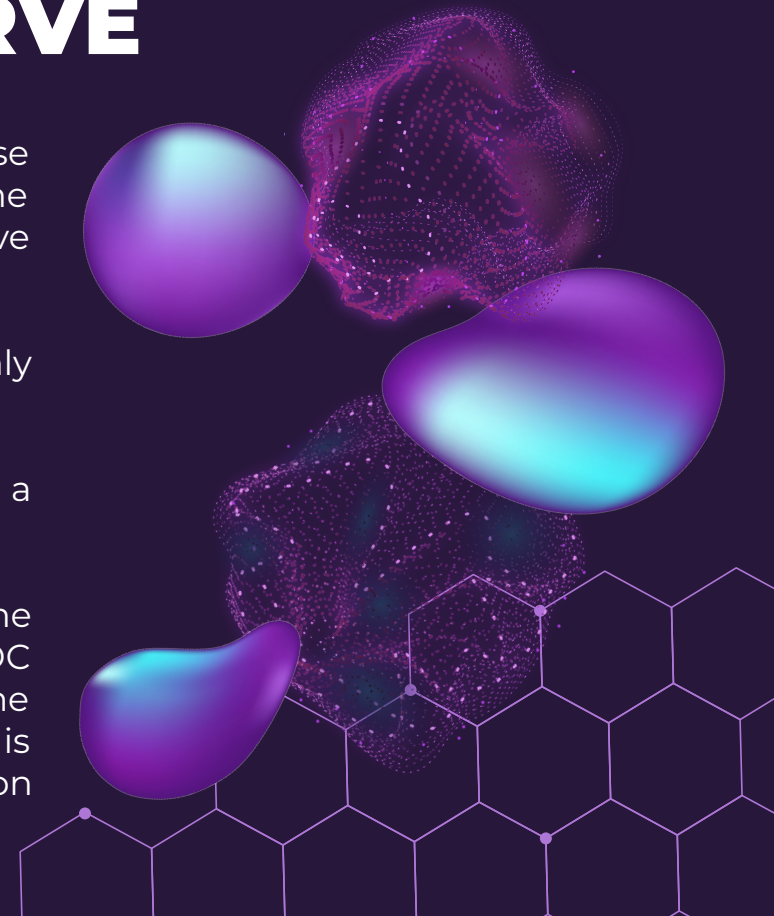
$$Precision = \frac{TP}{TP + FP}$$

ROC AUC CURVE

ROC AUC Curve is graphical plot is made by plotting the False positive rate on the x-axis and the True Positive Rate on the y-axis. The uses of the ROC curves and precision/recall curve are as follows:

- ROC curves should be used when there are roughly equal numbers of observations for each class.
- Precision-Recall curves should be used when there is a moderate to large class imbalance.

The main reason for this optimistic picture is because of the use of true negatives in the False Positive Rate in the ROC Curve and the careful avoidance of this rate in the Precision-Recall curve. The reason for this recommendation is that ROC curves present an optimistic picture of the model on datasets with a class imbalance.



FUTURE TRENDS

There is work ongoing for binary class imbalance problems but only a few on multi-class imbalance problems. The reasons are that binary class issues are more common as well as the complexity associated is lesser.

Studies have shown that class overlapping hinders classification more than imbalance. So, preprocessing with feature selection and engineering is more important.

In the future, the rapid development of big data computing most probably will shape the way classification tasks are performed and with anomaly patterns existing in most real-world problems, class imbalance problem is inevitable.

WORK DISTRIBUTION

Ayush Agrawal

Data level approach (Undersampling & Oversampling), z-SVM, Measures of Classification

Swetha Vipparla

Data level approach(Feature Selection, One-Class learning, Cost-sensitive learning

Shubh Agarwal

Challenges with Class imbalance classification, KNN with exemplar-based generalization, Measures of Classification

Rohan Madineni

Challenges with Class imbalance classification, KNN with exemplar-based generalization, Ensemble Methods

THANK YOU!