Advanced Techniques for Mitigating Abnormal Instances and Class Imbalance in High-Dimensional Data Classification

Surani Matharaarachchi,
PhD in Statistics

Presented at 2025 SSC Annual Meeting in Saskatoon

University of Manitoba

Machine Learning for Complex Data

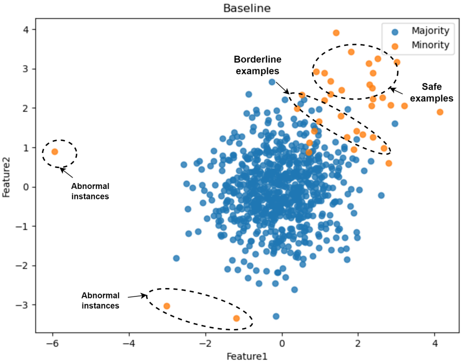
- The rapid advancement of science and technology has resulted in increasingly complex datasets
- Predictive Modeling
- Make data-driven decisions
- Challenges in Predictive Modeling
    - Class Imbalance Issue
    - Abnormal Instances
    - Curse of Dimensionality
    - Categorical Features

RQ 1

Class Imbalance Issue

Class Imbalance Issue

- Occurs when the number of instances in different classes is significantly disproportionate.
- Issue: Leads to biased models and decreases predictive accuracy.
- Abnormal Instances



Synthetic Minority Oversampling Technique (SMOTE)

- Resampling
- Balancing the Dataset:
  - Create new samples for the minority class.
- Technique:
  - Interpolate between randomly chosen minority class samples and their nearest neighbors.
  - $p_{new} = p_0 + \alpha(p_3 - p_0)$

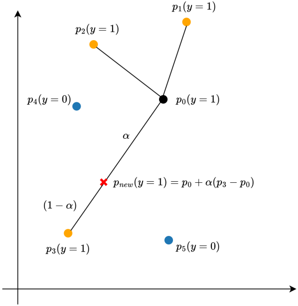


Figure: SMOTE data generation

Limitation with SMOTE

- Challenged by outliers within the minority class.

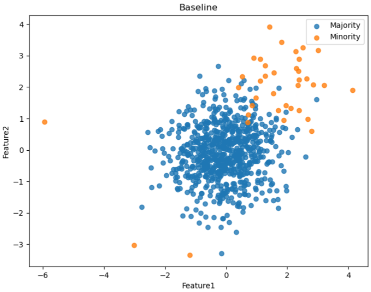


Figure: Original Data

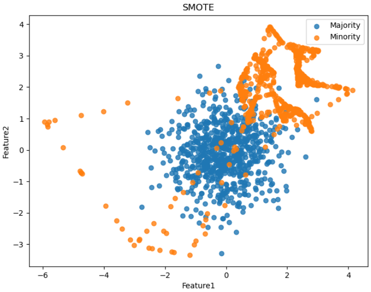


Figure: Re-sampled data with SMOTE

Proposed Solution [7]

- Technique:
  - Use a weighted average of neighbouring instances.
  - $p_{new} = \frac{\sum_{j=1}^{k}(w_j \times p_j)}{\sum_{j=1}^{k} w_j}, j = 1, \ldots, k$
  - Improve robustness against outliers and noisy data.
  - Learn from a more extensive set of nearest neighbours.
- Challenge:
  - Selecting suitable weights to enhance resilience to outliers and noisy data.

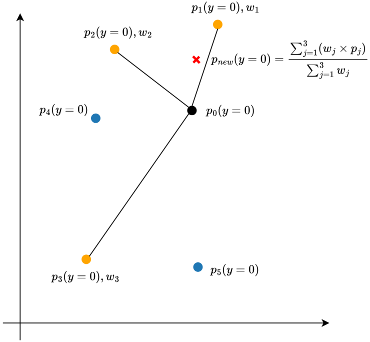


Figure: Proposed method data generation

How to Define Weights?

- Distance-based approach: Higher weights for closer instances in feature space.
- Use inverse distance to the median centroid of the minority class.
- Developing new SMOTE extensions:

1. Distance extSMOTE
2. Dirichlet extSMOTE [1]
3. FCRP SMOTE - SMOTE with Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model

Handling imbalanced data with abnormal minority instances

Machine Learning with Applications 18 (2024) 100597

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

ELSEVIER

Enhancing SMOTE for imbalanced data with abnormal minority instances

Surani Matharaarachchi [a,*], Mike Domaratzki [b], Saman Muthukumarana [a]

[a] Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[b] Department of Computer Science, Western University, London, ON, N6A 5B7, Canada

1. Distance extSMOTE

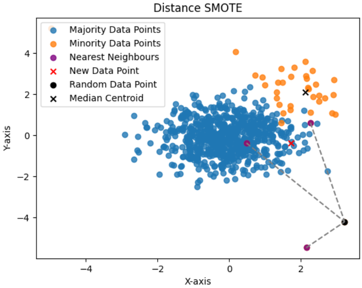
- $d_j \in \mathbb{R}$ is the Euclidean distance between the median centroid of the minority class and the nearest neighbours
- $w_j = d_{j,norm}^{-1}$ = Normalized inverse distance

Algorithm Distance ExtSMOTE

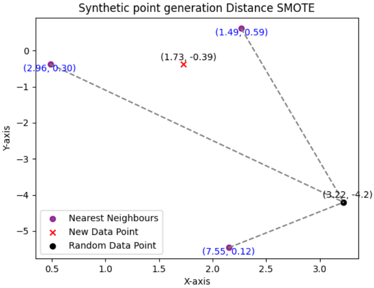
Require: $X \in \mathbb{R}^{n \times p}$ the features, $Y \in \{0,1\}^n$ the binary class label outputs.
Require: $k \in \mathbb{N}$ the number of neighbors to select for the k-Nearest Neighbors.
Ensure: Generated data $X_{new} \in \mathbb{R}^{q \times p}$ and $Y_{new} \in \{0,1\}^q$ with q points created.

Denote by S_1 the number of points labelled as the minority class and S_0 the number of points labelled as the majority class.
Initialize X_{new} and Y_{new} as empty vectors.
Obtain the median centroid (μ) of the minority class.
while $S_1 < S_0$ **do**
    Filter $\mathcal{D} = \{X_i | Y_i = 1\}$, the set of points labeled as minority class 1.
    Randomly choose $r \in \mathcal{D}$ and find the indices of its k nearest neighbors, $r_1, \dots, r_k$.
    Consider the inverse distances, from μ, to each nearest neighbour as weights, $w_j = d_j^{-1}$
    $x^{new} \leftarrow \frac{\sum (w_j \times x_{r_j})}{\sum w_j}$ for all j from 1 to k.
    $y^{new} \leftarrow 1$
    $S_1 = S_1 + 1$
    Append x^{new} to X_{new}, append y^{new} to Y_{new}
end while
return X_{new}, Y_{new}

1. Distance extSMOTE



(a) This scenario occurs when an outlier is chosen as a neighbouring point.



(b) The values within parentheses indicate (d_j, w_j).

Figure: An example of creating a sample - Distance extSMOTE

2. Dirichlet extSMOTE

- The Dirichlet distribution is defined by a set of parameters $\alpha_1, \alpha_2, \ldots, \alpha_K$, where K is the dimensionality of the probability simplex.
- The distribution is parameterized by $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]$, which can be considered pseudo-counts or prior observations.
- Let $p = [p_1, p_2, \ldots, p_k]$ be a K-dimensional vector such that for all $j = 1, \ldots, k$ we have $p_j \geq 0, j = 1, 2, \ldots, k$ and $\sum_{j=1}^{K} p_j = 1$.
- The pdf of the Dirichlet distribution for a point p on the simplex [2]:

$$w_j = P(p|\alpha) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_K) \stackrel{\text{def}}{=} \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \quad (1)$$

Dirichlet extSMOTE

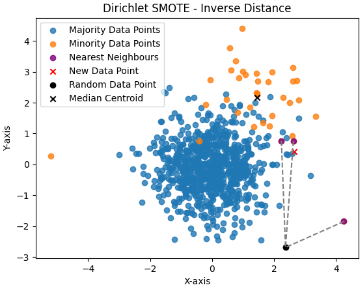
Algorithm Dirichlet ExtSMOTE

Require: $X \in \mathbb{R}^{n \times p}$ the features, $Y \in \{0,1\}^n$ the binary class label outputs, $k \in \mathbb{N}$ the number of neighbors to select for the k-Nearest Neighbors, $m > 0 \in \mathbb{R}$, the multiplier of the parameter of the distribution.

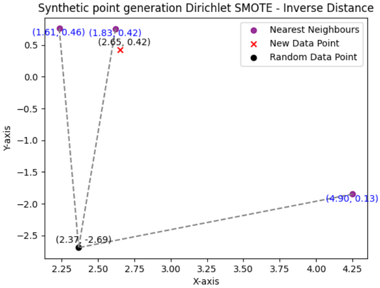
Ensure: Generated data $X_{new} \in \mathbb{R}^{q \times p}$ and $Y_{new} \in \{0,1\}^q$ with q points created.

```
Denote by S_1 the number of points labelled as the minority class and S_0 the number of points labelled as the majority class.
Initialize X_new and Y_new as empty vectors.
Obtain the median centroid (μ) of the minority cluster.
while S_1 < S_0 do
    Filter 𝒟 = {X_i | Y_i = 1}, the set of points labeled as minority class 1.
    Randomly choose r ∈ 𝒟 and find the indices of its k nearest neighbors, r_1, ..., r_k.
    if Type is 'Inverse distance (D)' then
        Calculate the distances, D = [d_1, ..., d_k] from μ to each nearest neighbour and obtain the reciprocal of each distance D^{-1} = [1/d_1, ..., 1/d_k]. Then α = D^{-1} × m
    else if Type is 'Uniform Vector (UV)' then
        Generate a vector α = 1_k × m, where 1_k = [1, ..., 1]
    else if Type is 'Uniform Distribution (UD)' then
        Generate vector U of size k from uniform(0, 1) distribution, then α = U × m.
    end if
    Use α as parameters to the Dirichlet Distribution and generate random weights w_j ~ Dir(α)
    x^new ← Σ w_j x_{r_j} for all j from 1 to k, as Σ w_j = 1
    y^new ← 1, S_1 = S_1 + 1
    Append x^new to X_new, append y^new to Y_new
end while
```

2. Dirichlet extSMOTE (Inverse Distance)



(a) This scenario occurs when an outlier is chosen as a neighbouring point.



(b) The values within parentheses indicate (d_j, w_j).

Figure: An example of creating a sample - Dirichlet extSMOTE

Synthetic Data Generation

- $X_{minority-outliers} \sim \mathcal{N}_{2\times 2}(\mu^{(1)}_{2\times 1}, \Sigma^{(1)}_{2\times 2})$
- $X_{majority} \sim \mathcal{N}_{2\times 2}(\mu^{(2)}_{2\times 1}, \Sigma^{(2)}_{2\times 2})$
- $X_{outliers} \sim \mathrm{Uniform}([-10, 10]^2)$

$$\mu^{(1)}_{2\times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}_{2\times 1}, \Sigma^{(1)}_{2\times 2} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}_{2\times 2}$$

$$\mu^{(2)}_{2\times 1} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times 1}, \Sigma^{(2)}_{2\times 2} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}_{2\times 2}$$

Synthetic Data Generation

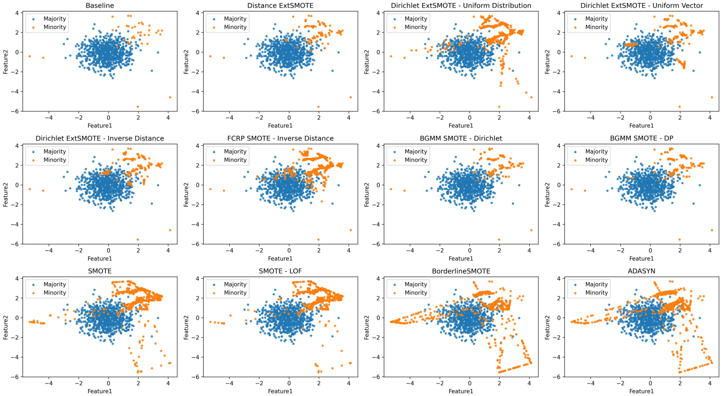


Figure: Comparison of resampled data

Synthetic Data Generation

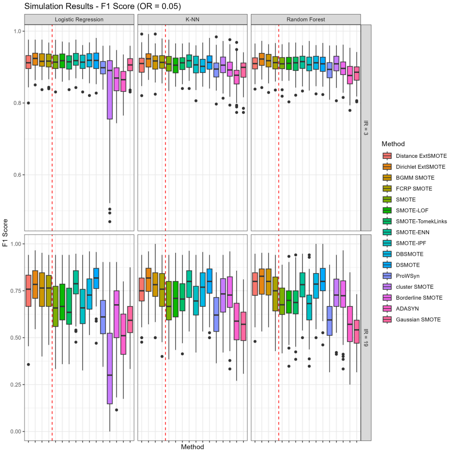


Figure: F1 Scores for 100 simulated datasets with 5-fold cross validation

Application Data

Table: Characteristics of the binary class datasets used in the computational study.

No	Dataset	Instances	Features	Minority class	Majority class	%Minority	%Majority	IR	Presence of LOF Outliers
1	yeast6	1484	8	EXC	Remaining classes	2.36	97.64	41.40	Yes
2	yeast5	1484	8	EXC, ERL	Remaining classes	2.70	97.30	36.10	Yes
3	yeast-1289vs7	947	8	VAC	NUC, CYT, ERL, POX	3.17	96.83	30.57	Yes
4	yeast4	1484	8	ME2	Remaining classes	3.44	96.56	28.10	Yes
5	yeast-2vs8	483	8	POX	CYT	4.14	95.86	23.15	Yes
6	glass12357vs6	214	9	6	Remaining classes	4.21	95.79	22.78	Yes
7	yeast-1458vs7	693	8	VAC	NUC, ME3, ME2, POX	4.33	95.67	22.10	Yes
8	oil	937	49	minority	majority	4.38	95.62	21.85	No
9	abalone9_18	731	7	9, 18	Remaining classes	5.75	94.25	16.40	Yes
10	glass12367vs5	214	9	5	Remaining classes	6.07	93.93	15.46	Yes
11	thyroid_sick	3772	52	sick	healthy	6.12	93.88	15.33	Yes
12	yeast-1vs7	459	8	VAC	NUC	6.54	93.46	14.30	Yes
13	us_crime	1994	100	>0.65	$<=0.65$	7.52	92.48	12.29	Yes
14	glass12vs5	159	9	5	1, 2	8.18	91.82	11.23	Yes
15	spectrometer	531	93	$>=44$	<44	8.47	91.53	10.80	Yes
16	landsat_satellite	6435	36	2	Remaining classes	9.73	90.27	9.28	Yes
17	mfeatmor0	2000	6	0, 1	Remaining classes	10.00	90.00	9.00	Yes
18	yeast3	1484	8	ME3	Remaining classes	10.98	89.02	8.10	Yes
19	mfeatmor01	2000	6	0	Remaining classes	20.00	80.00	4.00	Yes
20	glass123vs567	214	9	5, 6, 7	Remaining classes	23.83	76.17	3.20	Yes
21	parkinsons	195	22	1	0	24.62	75.38	3.06	Yes
22	habermans_survival	306	3	2	1	26.47	73.53	2.78	Yes
23	glass23567vs1	214	9	1	Remaining classes	32.71	67.29	2.06	Yes
24	breast_cancer	569	30	M	B	37.26	62.74	1.68	Yes
25	banknote	1372	4	1	Remaining classes	44.46	55.54	1.25	Yes

Application Results

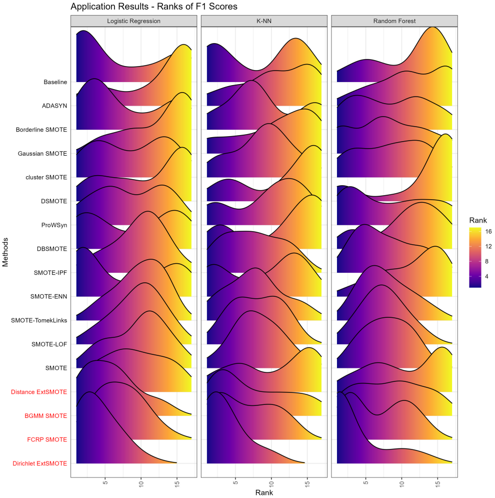


Figure: F1 Score Ranks for the datasets with 100 $\times$ 5-fold cross validation

RQ 2

High-Dimensional Data and Presence of Categorical Data

Matharaarachchi, S., Domaratzki, M. and Muthukumarana, S. (2025). Deep-ExtSMOTE: Integrating Autoencoders for Advanced Mitigation of Class Imbalance in High-Dimensional Data Classification. The Manuscript is submitted for the Journal of Big Data Research.

High-Dimensional Data

- Curse of Dimensionality
  - A large number of features relative to the available data, "large p, small n" problem [4].
  - Challenges:
    - Data Sparsity
    - Increased Model Complexity and Overfitting
    - Computational Challenges
- Feature Reduction
  - A critical strategy to address the challenges of high dimensionality in class imbalance [3, 5, 6].

Autoencoders

- Neural network models are designed to learn a compressed, lower-dimensional input data representation.
- Composed of two parts:
  - **Encoder:** Takes the input data and maps it to a lower-dimensional representation.
  - **Decoder:** Takes the low-dimensional representation produced by the encoder and reconstructs the original input data.

Encoder

- The encoder consists of several layers that sequentially reduce the dimensionality of the input $\mathbf{x}^{(i)}$, where i denotes the i^{th} observation.
- The transformation at each layer can be mathematically expressed as:

$$
\begin{aligned}
\mathbf{h}_1^{(i)} &= f_1(\mathbf{W}_1^T \cdot \mathbf{x}^{(i)} + \mathbf{b}_1) \\
\mathbf{h}_2^{(i)} &= f_2(\mathbf{W}_2^T \cdot \mathbf{h}_1^{(i)} + \mathbf{b}_2) \\
&\dots \\
\mathbf{z}^{(i)} &= f_n(\mathbf{W}_n^T \cdot \mathbf{h}_{n-1}^{(i)} + \mathbf{b}_n),
\end{aligned}
$$

where $\mathbf{h}_1^{(i)}, \mathbf{h}_2^{(i)}, \dots$ are the intermediate representations at each layer, $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_n$ are the weight matrices for each layer, $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ are the biased vectors for each layer, $\mathbf{z}^{(i)}$ is the latent vector, representing the lower-dimensional feature space and $f_1, f_2, \dots, f_n$ represent the non-linear activation functions used in each layer.

Decoder

- The decoder takes the latent vector $\mathbf{z}^{(i)}$ and maps it back to a reconstruction of the original input.
- The process at each layer of the decoder can be expressed as:

$$\begin{aligned}
\mathbf{h'}^{(i)}_{n-1} &= f'_{n-1}(\mathbf{W'}^{T}_{n-1} \cdot \mathbf{z}^{(i)} + \mathbf{b'}_{n-1}) \\
\mathbf{h'}^{(i)}_{n-2} &= f'_{n-2}(\mathbf{W'}^{T}_{n-2} \cdot \mathbf{h'}^{(i)}_{n-1} + \mathbf{b'}_{n-2}) \\
&\ \ \cdots \\
\hat{\mathbf{x}}^{(i)} &= f'_{1}(\mathbf{W'}^{T}_{1} \cdot \mathbf{h'}^{(i)}_{2} + \mathbf{b'}_{1}),
\end{aligned}$$

where $\hat{\mathbf{x}}^{(i)}$ is the constructed output for the i^{th} observation, which approximates $\mathbf{x}^{(i)}$. $\mathbf{h'}^{(i)}_{n-1}, \mathbf{h'}^{(i)}_{n-2}, \ldots$ are the intermediate representations in the decoder. $\mathbf{W'}_1, \mathbf{W'}_2, \ldots, \mathbf{W'}_{n-1}$ are the weight matrices for each decoder layer, where $\mathbf{b'}_1, \mathbf{b'}_2, \ldots, \mathbf{b'}_{n-1}$ are the biased vectors for each decoder layer. $\mathbf{z'}^{(i)}$ is the latent vector, representing the lower-dimensional feature space and $f'_1, f'_2, \ldots, f'_{n-1}$ represent the non-linear activation functions used in each decoder layer.

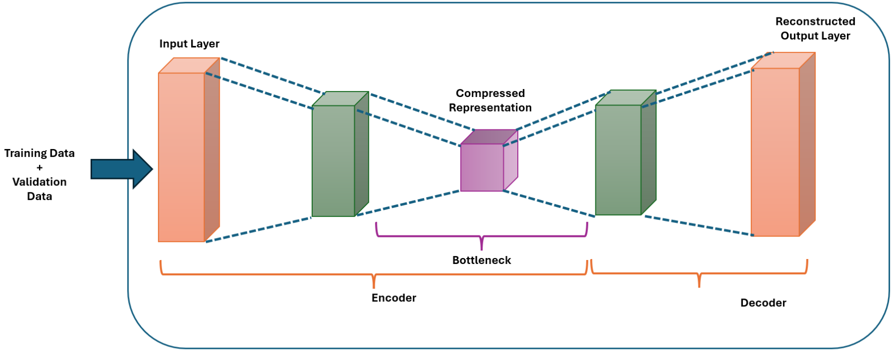
Train Multi-layer Autoencoder

- To train the multi-layer autoencoder, we use a loss function that measures the difference between the input $\mathbf{x}^{(i)}$ and the reconstructed output $\hat{\mathbf{x}}^{(i)}$. The most common loss function for this purpose is Mean Squared Error (MSE):

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}||^2 \tag{2}$$

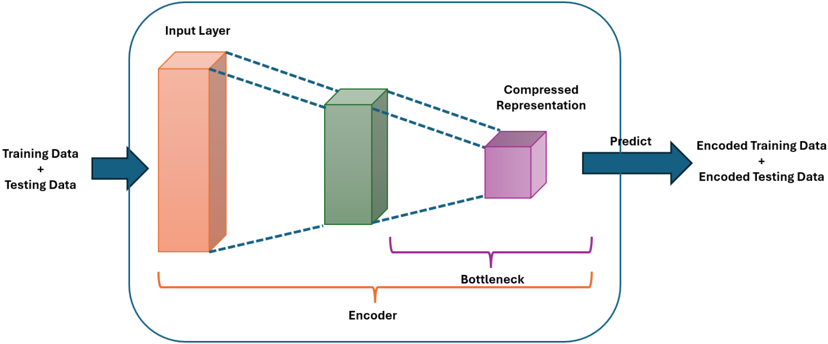
5. Deep-ExtSMOTE

- Autoencoder + Dirichlet ExtSMOTE
- Step 1: Train the Autoencoder



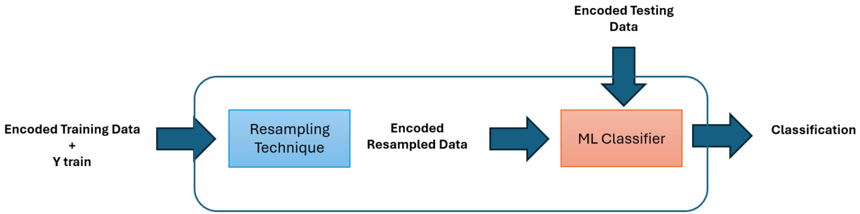
5. Deep-ExtSMOTE

- Step 2: Extract Encoded Representation



5. Deep-ExtSMOTE

- Step 3: Resampling and Classification



Simulation Results

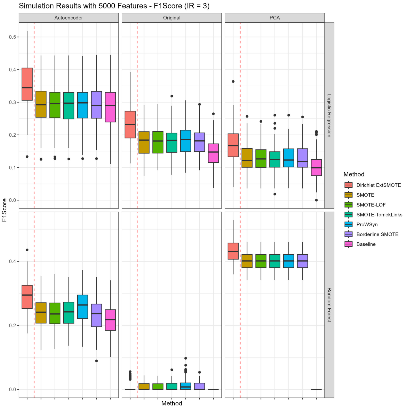


Figure: F1-Score distribution for 100 trials using simulated datasets with 1000 samples and 5000 features (2000 informative), with an imbalance ratio (IR) of 3.

Application Results

- Application 1: Isolet (Isolated Letter Speech Recognition) - Continuous Binary Classification
  - **Objective:** Accurately recognize spoken letters based on high-dimensional acoustic features extracted from voice recordings of multiple speakers.
  - The dataset includes 617 continuous features, representing processed characteristics of the audio signals.
  - Scenario 1: Original Isolet Dataset
    - Dataset includes 7797 samples, resulting in a feature-to-sample ratio of approximately 0.0791.
  - Scenario 2: Reduced Isolet Dataset
    - Selected a subset of 1000 samples from the original 7797 samples. This adjustment resulted in a feature-to-sample ratio of 0.617.

Application Results

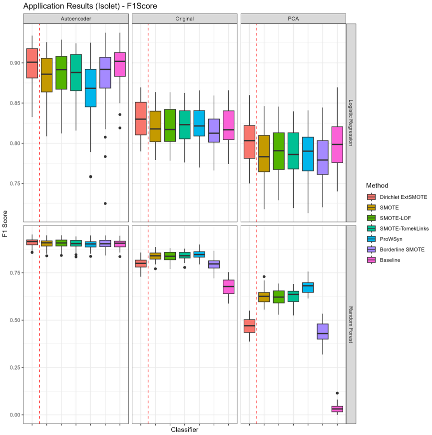


Figure: F1 Scores for the Isolet dataset across 50 training and test splits.

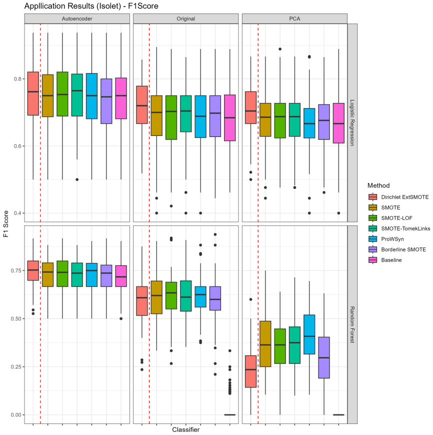


Figure: F1 Scores for the reduced Isolet dataset across 50 training and test splits.

Application Results

- Application 2: Chile (Categorical Binary Classification)

  - **Objective:** Predict the yield of 204 chile pepper genotypes from multi-environment trials in New Mexico, USA.

  - Conduct experiment by starting with 2,500 features and increasing the number of features to 7,500.

  - Feature-to-sample ratio ranging from approximately 12.25 to 37.7.

Application Results

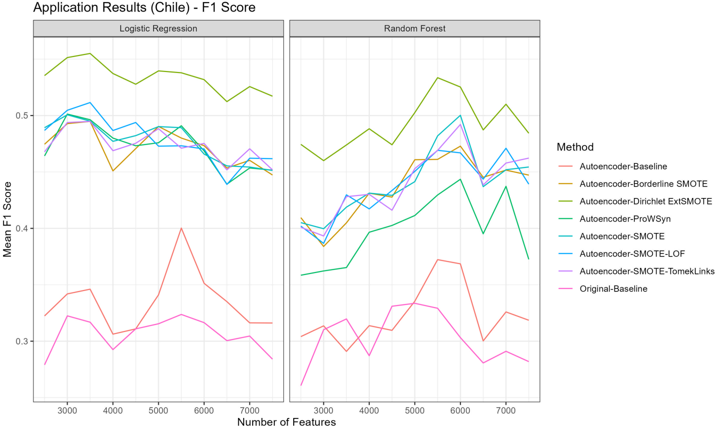


Figure: F1 score comparison with varying feature numbers.

Application Results

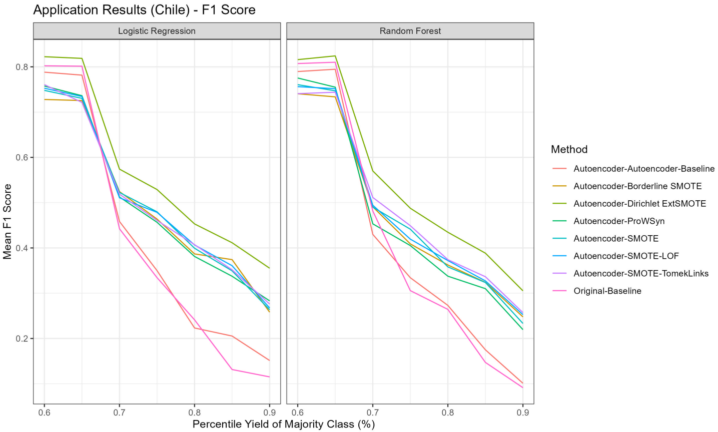


Figure: F1 score comparison with varying imbalance ratios.

References

[1] Bej, S., N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer (2021). Loras: an oversampling approach for imbalanced datasets. *Machine learning 110*(2), 279–301.

[2] Bela, A, F., K. Amol, and G. Maya, R (2010). Introduction to the dirichlet distribution and related processes. Technical report, University of Washington Department of Electrical Engineering.

[3] Garzon, M. (2022). *Dimensionality reduction in data science*. Cham, Switzerland: Springer.

[4] Huynh, P.-H., V. H. Nguyen, and T.-N. Do (2020). Improvements in the large p, small n classification issue. *SN computer science 1*(4), 207–.

[5] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2021). Assessing feature selection method performance with class imbalance data. *Machine learning with applications 6*, 100170–.

[6] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2022). Minimizing features while maintaining performance in data classification problems. *PeerJ. Computer science 8*, e1081–e1081.

[7] Matharaarachchi, S., M. Domaratzki, and S. Muthukumarana (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*.

Acknowledgment

University of Manitoba
Manitoba Centre for Health Policy
Data | Insight | Informing Solutions

UM
Faculty of Graduate Studies

Digital Research Alliance of Canada

Winnipeg

Manitoba

Thank You!

Contact: matharas@myumanitoba.ca
Personal Website: https://suranimatharaarachchi.com