

Tabla de Contenido

1. Introducción	1
1.1. Motivación	1
1.2. Contexto	2
1.3. Objetivo general	3
1.4. Objetivos específicos	3
1.5. Alcances	4
2. Antecedentes	5
2.1. Preprocesamiento	5
2.2. Visualización	5
2.3. Enfoques actuales al problema de datos desbalanceados	6
2.4. Resamplio de datos	7
2.4.1. Oversampling	8
2.4.2. Undersampling	10
2.4.3. Combinación oversampling y undersampling	12
2.5. Ensemble learning	13
2.5.1. Métodos tipo bagging	13
2.5.2. Métodos tipo boosting	15
2.6. Redes neuronales	16
2.7. Métricas de evaluación	19
3. Metodología	21
3.1. Montaje experimental	21
3.2. Caracterización de datos y falla de entrenamiento	22
3.3. Preprocesamiento	22
3.4. Definición etiquetas y conjuntos	24
3.5. Selección y desarrollo de algoritmos	26
3.6. Visualización gráfica de métodos de resamplio	28
3.7. Selección de hiperparámetros	28
3.8. Evaluación	29
4. Resultados y discusión	30
4.1. Caracterización de datos y falla de entrenamiento	30

4.2.	Preprocesamiento	34
4.3.	Visualización gráfica de clases en el caso base	36
4.4.	Evaluación del desempeño de los modelos	37
4.4.1.	Resumen de los resultados de evaluación	37
4.4.2.	Modelo caso base	41
4.4.3.	Modelo Random Over Sampler + RNM	42
4.4.4.	Modelo SMOTE + RNM	44
4.4.5.	Modelo All-KNN + RNM	45
4.4.6.	Modelo Neighbourhood Cleaning Rule + RNM	47
4.4.7.	Modelo SMOTE-ENN + RNM	48
4.4.8.	Modelo Balanced Bagging Classifier	49
4.4.9.	Modelo Easy Ensemble Classifier	50
4.5.	Evaluación capacidad de generalización	51
4.5.1.	Detección de falla de entrenamiento	51
4.5.2.	Detección de falla nueva	52
5.	Conclusiones y trabajos futuros	55
5.1.	Trabajos futuros	56
	Bibliografía	57
	Anexos	61
	Anexo A. Resultados modelos restantes	61
A.1.	Modelo ADASYN + RNM	61
A.2.	Modelo Borderline SMOTE + RNM	62
A.3.	Modelo SVMSMOTE + RNM	63
A.4.	Modelo Cluster Centroids + RNM	64
A.5.	Modelo Condensed Nearest Neighbour + RNM	65
A.6.	Modelo Edited Nearest Neighbours + RNM	66
A.7.	Modelo Repeated Edited Nearest Neighbours + RNM	67
A.8.	Modelo Near Miss + RNM	68
A.9.	Modelo One Sided Selection + RNM	69
A.10.	Modelo Random Under Sampler + RNM	70
A.11.	Modelo Tomek Links + RNM	71
A.12.	Modelo SMOTE + Tomek Links + RNM	72
A.13.	Modelo Balanced Random Forest Classifier	73
A.14.	Modelo RUS Boost Classifier	73
	Anexo B. Códigos	74
B.1.	Gráficos de dispersión	74
B.2.	Definición de conjuntos de entrenamiento y evaluación	77

B.3. Ajuste de hiperparámetros	78
B.4. Entrenamiento y evaluación de modelos	79