Tabla de Contenido

1. Introducción **1**
- 1.1. Identificación y formulación del problema . . . 2
- 1.2. Objetivos . . . 6
  - 1.2.1. Objetivo general . . . 6
  - 1.2.2. Objetivos específicos . . . 6
- 1.3. Organización del trabajo de memoria . . . 6

2. Marco Teórico y Estado del Arte **8**
- 2.1. Modelos predictivos . . . 8
  - 2.1.1. Balanced Random Forest . . . 8
  - 2.1.2. Extreme Gradient Boosting . . . 8
- 2.2. Factores que dificultan el aprendizaje . . . 11
- 2.3. Técnicas de balanceo de datos . . . 13
  - 2.3.1. Sobremuestreo . . . 13
  - 2.3.2. Submuestreo . . . 18
  - 2.3.3. Combinación de aumento y limpieza de datos . . . 20
- 2.4. Técnicas de balanceo a nivel algoritmo . . . 22
- 2.5. Evaluación . . . 24

3. Metodología **28**

4. Experimentos y resultados **33**
- 4.1. Replica del clasificador de curvas de luz de ALeRCE . . . 33
- 4.2. Entrenamiento de XGBoost con selección de hiperparámetros . . . 37
- 4.3. Entrenamiento de XGBoost en conjunto con primeras técnicas de balance de datos . . . 42
- 4.4. Entrenamiento con técnicas de Sobremuestreo . . . 53
- 4.5. Entrenamiento con técnicas de Submuestreo y Sobremuestreo . . . 59
- 4.6. Entrenamiento con técnicas híbridas . . . 66
- 4.7. Entrenamiento con combinación de técnicas de balance a nivel de datos y Cost Sensitive Learning . . . 69
- 4.8. Entrenamiento con Focal Loss Cross-Entropy como función de pérdida . . . 74

5. Análisis Estadístico de Resultados **82**

6. Evaluación de mejores modelos con datos nuevos de ALeRCE **87**

7. Conclusiones **96**

Bibliografía **99**

Anexos **102**

A. Matrices de confusión obtenidas por ALeRCE en [4] **103**

B. Glosario hiperparámetros **105**
B.1. XGBoost 105
B.2. Técnicas de *Oversampling* 105
B.3. Técnicas de *Undersampling* 106

C. Cálculo de gradientes y Hessiano **107**
C.1. Gradiente de la función *softmax* 107
C.2. Gradiente de la función Focal Loss Cross-Entropy 108
C.3. Hessiano de la función Focal Loss Cross-Entropy 109