An ML Approach to Optimizing Project Proposals for ALMA Observatory

Arnav Boppudi, Ryan Lipps, Noah McIntire, Kaleigh O’Hara, Brendan Puglisi, Antonios Mamalakis

UVA | SCHOOL of DATA SCIENCE

Introduction:

Every year, astronomers around the world submit research proposals to the Atacama Large Millimeter Array (ALMA), the largest radio telescope array in the world. The aim of this work is to streamline the proposal process for astronomers submitting projects to ALMA by suggesting technical setups relevant to their research. Our work provides researchers insight and assistance on where to set measurements in the project technical plan when submitting, with the intent of optimizing the insight they can extract from the data. We introduce a pipeline of supervised and unsupervised machine learning models, each using various representations of the title and abstract of an incoming proposal. This creates an interpretable series of models that can assist in recommending frequency ranges in project setups. The full dataset consists of 4,586 accepted projects, each with a title and abstract, accounting for 67,439 associated measurements as model inputs.

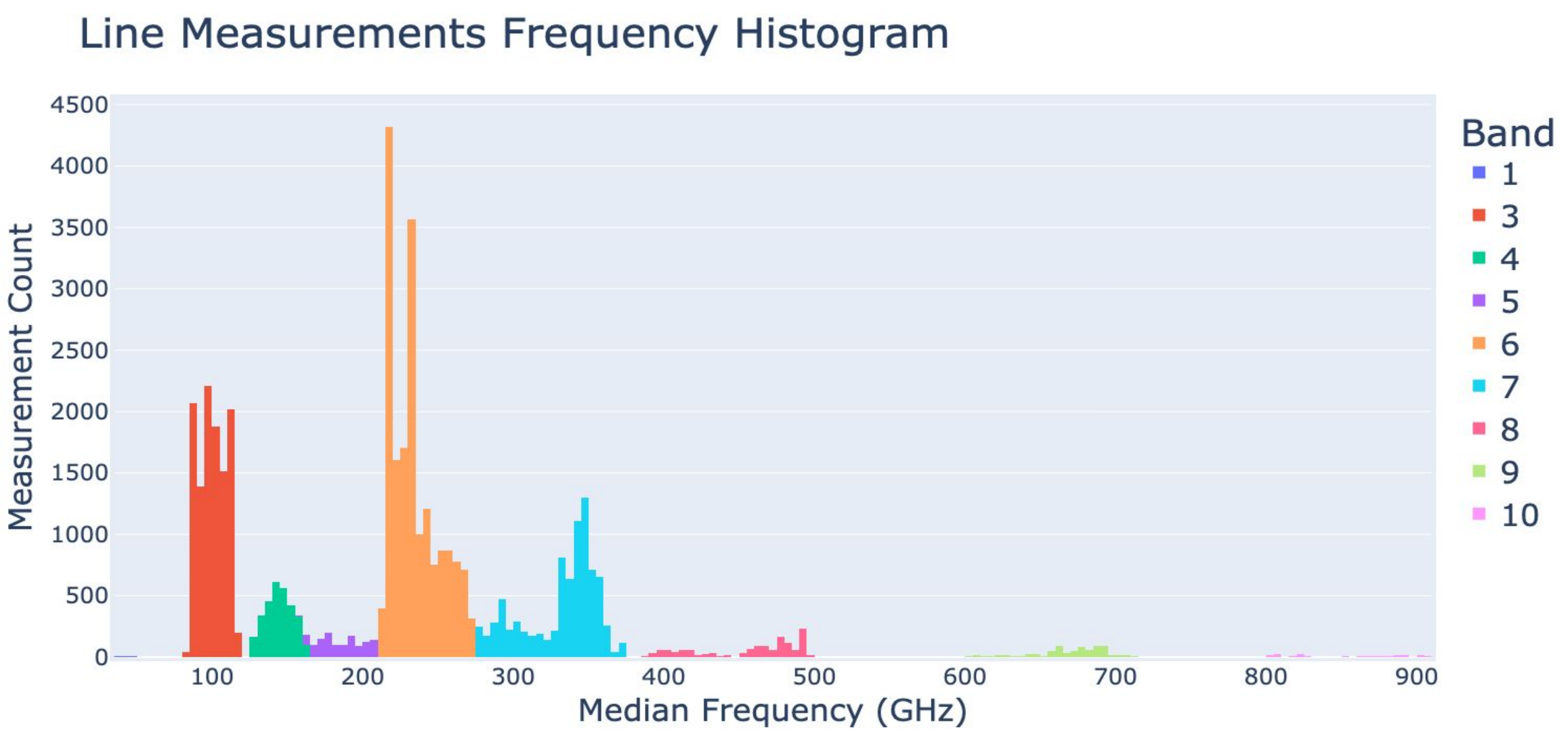


Figure 1: This histogram showcases the line projects measurement count distribution across all the frequency bands (from 35GHz to 950GHz at two-decimal-point precision). It is important to note that these 44,232 measurements are sourced from the 3,628 accepted projects that are labeled as line projects.

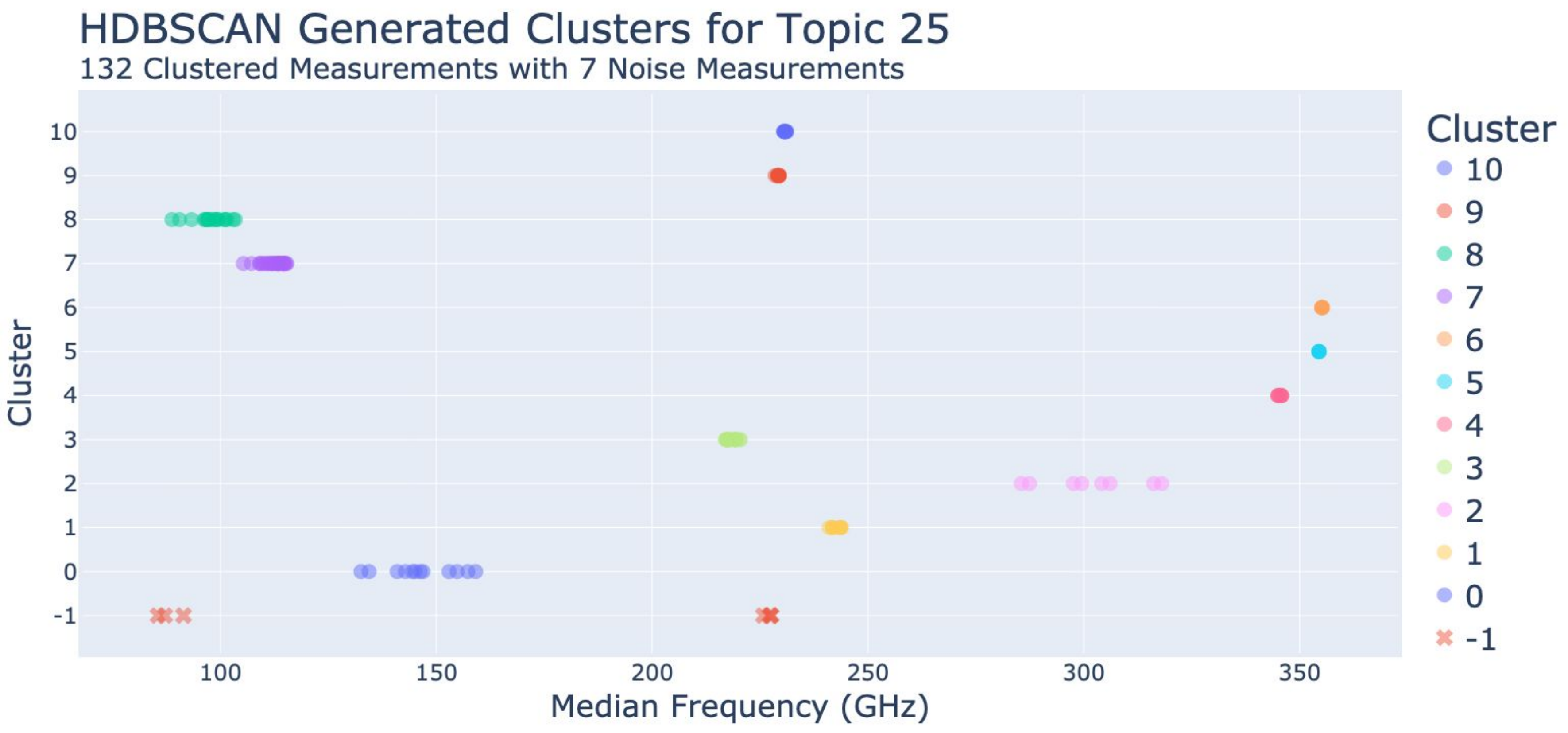


Figure 2: Inspection of HDBSCAN Clusters for Topic 25. 18 training projects are assigned to topic 25 from the LDA model, accounting for 139 measurements. HDBSCAN produces 12 clusters, with 7 measurements (0.5% of measurements within the cluster) identified as noise (-1). The HDBSCAN algorithm differentiates the densities of similar-sized clusters across varied frequency ranges, such as the 8-measurement clusters 1 (241-244 GHz), 2 (286-318 GHz), and 4 (241.00-243.87GHz).

Method	Results
Logistic Regression	**90.02%** Overall Accuracy **96.41%** Line Prediction Accuracy **59.42%** Continuum Prediction Accuracy
HDBSCAN Clustering with Topics	**88.81%** of projects have at least one correct measurement **60.00%** of measurements in a project captured in the prediction
Band Prediction (weighted)	**69.70%** of the time the predicted set of 2 bands includes the correct band **100%** of bands predicted at least once
Band Prediction (unweighted)	**73.55%** of the time the predicted set of 2 bands includes the correct band **44.44%** of bands predicted at least once
Combined (weighted)	**67.17%** of projects have at least one correct measurement **44.72%** of measurements in a project captured in the prediction

Table 1: Performance of different models in the proposed pipeline

Model Prediction Pipeline

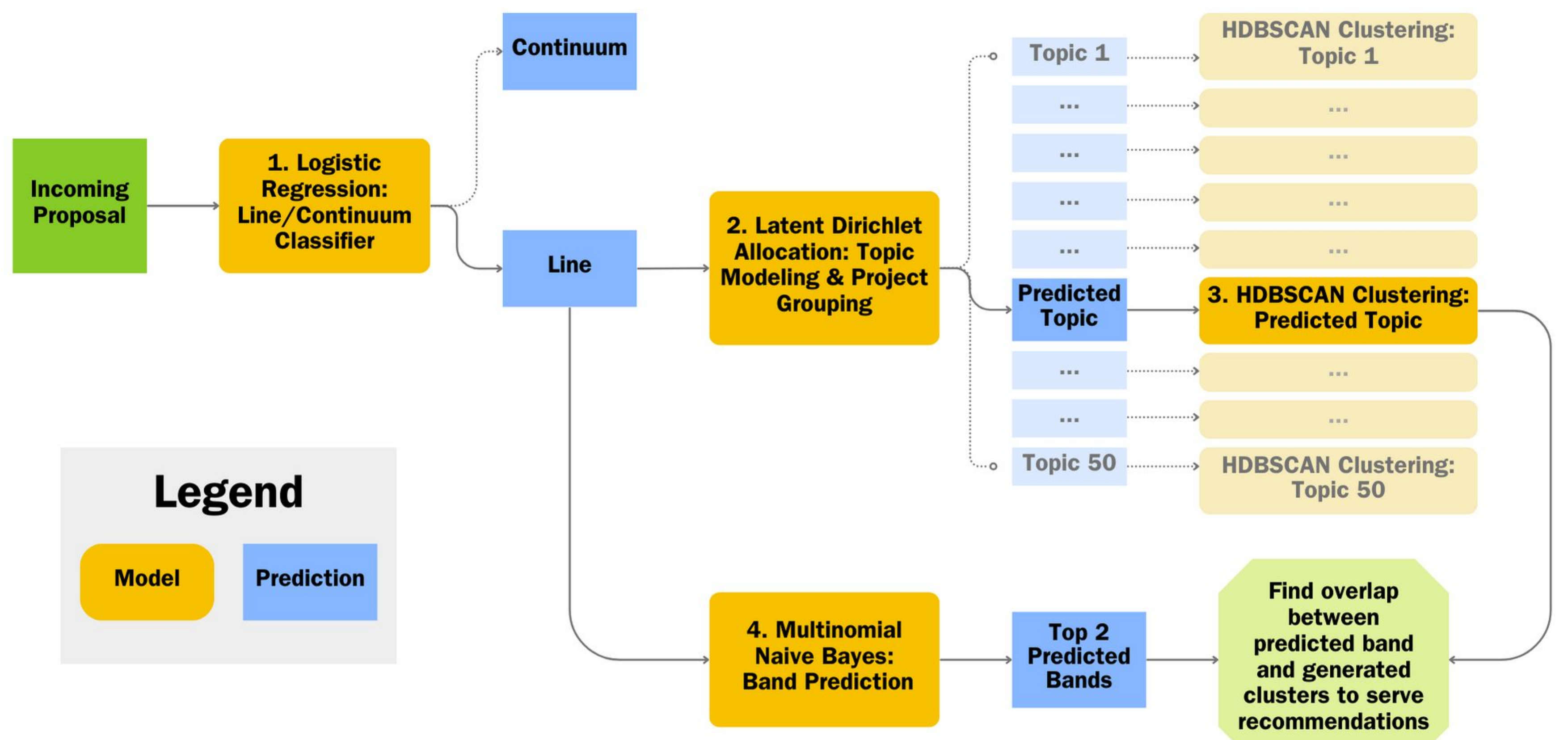

Incoming Proposal
1. Logistic Regression: Line/Continuum Classifier
Continuum
Line
2. Latent Dirichlet Allocation: Topic Modeling & Project Grouping
Topic 1
...
Predicted Topic
Topic 50
HDBSCAN Clustering: Topic 1
3. HDBSCAN Clustering: Predicted Topic
HDBSCAN Clustering: Topic 50
4. Multinomial Naive Bayes: Band Prediction
Top 2 Predicted Bands
Find overlap between predicted band and generated clusters to serve recommendations
Legend
Model
Prediction

Figure 3: Flowchart of process for incoming proposals. Numbers in model blocks reference steps below

1. Logistic Regression:

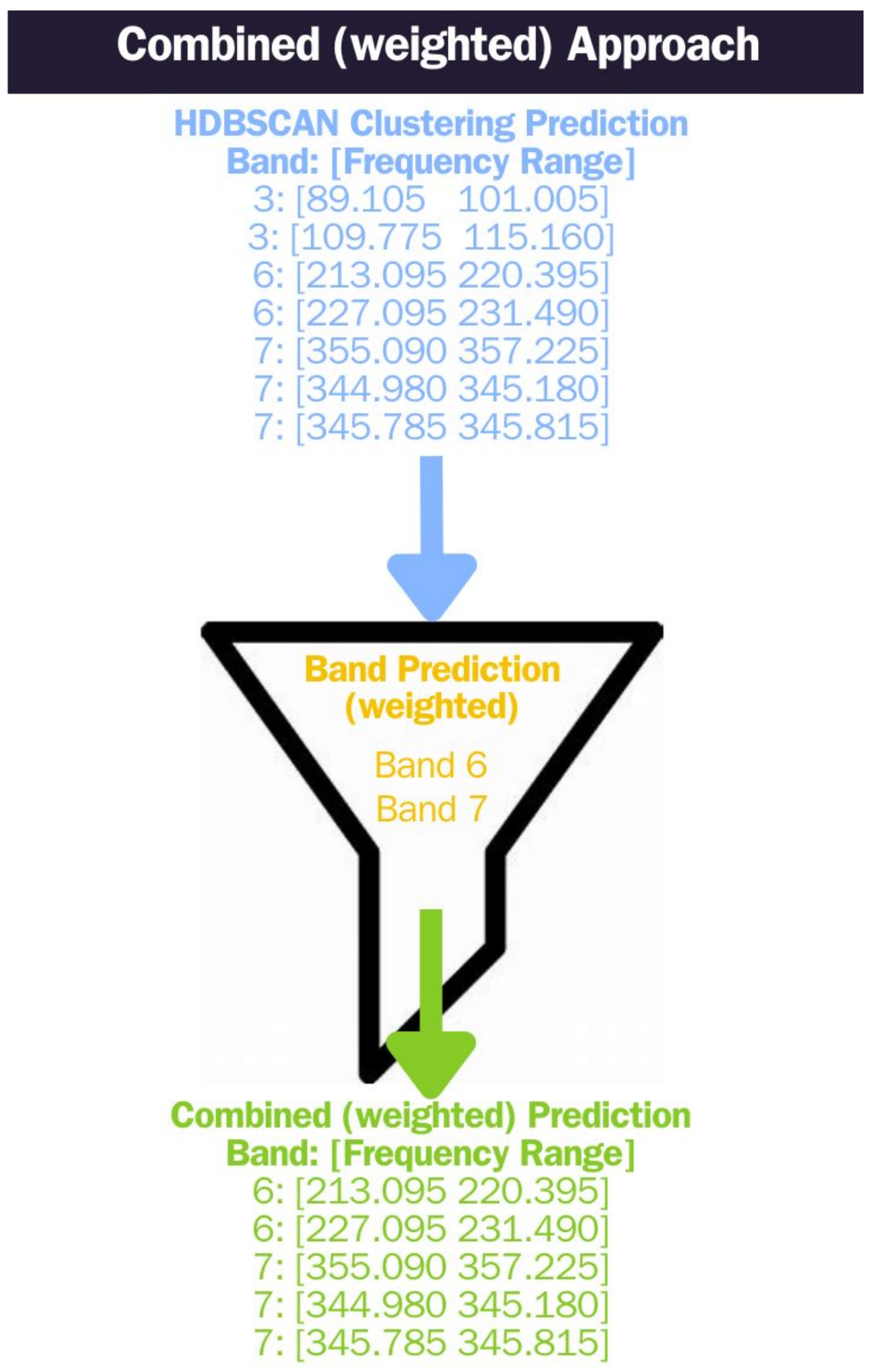
Purpose: Classify projects as either “line” or “continuum” based on textual descriptions

Method: Utilizes a binary logistic regression model with TF-IDF vectorization of text. A project is classified as “line” if the predicted probability exceeds 0.5.

2. LDA Topic Modeling:

Purpose: Group similar projects into topics

Method: Projects are assigned to topics for which they show the highest topic weight, based on the probabilistic mixture of words. We parameterized to generate 50 distinct topics.



Combined (weighted) Approach

HDBSCAN Clustering Prediction
Band: [Frequency Range]
3: [89.105 101.005]
3: [109.775 115.160]
6: [213.095 220.395]
6: [227.095 231.490]
7: [355.090 357.225]
7: [344.980 345.180]
7: [345.785 345.815]

Band Prediction (weighted)
Band 6
Band 7

Combined (weighted) Prediction
Band: [Frequency Range]
6: [213.095 220.395]
6: [227.095 231.490]
7: [355.090 357.225]
7: [344.980 345.180]
7: [345.785 345.815]

Figure 4: The Combined (weighted) approach filters the HDBSCAN Clustering Prediction with the Band Prediction ((weighted)

Future Work:

1) Compare performance and possibly replace Logistic Regression and Band Classification models with BERT models.

2) Finetune weights for Band Classification (weighted) approach for our specific imbalanced dataset.

3) Consult subject matter experts and fine tune text processing to ensure the LDA topics are both salient and discriminant.

3. HDBSCAN Clustering within Topics:

Purpose: Mine measurements within topics by identifying dense regions of measurements

Method: Extends the DBSCAN algorithm by dynamically adjusting the ϵ-parameter based on the density of data points. Accounting for varying densities generates clusters that are representative of the topic’s measurement distribution while providing focus on specific areas.

4. Band Classification - Multinomial Naive Bayes:

Purpose: Predict the most likely frequency bands a project pertains to, based on its textual description.

Method: Uses TF-IDF to vectorize text from titles and abstracts. Incorporates prior probabilities to address the band imbalance in measurement distribution. Features a weighted approach that adjusts priors based on band frequency to enhance precision and an unweighted approach that treats all bands equally.

Conclusion:

The Combined (weighted) process is useful to researchers submitting projects to ALMA and also proposal reviewers. It offers a way to gain valuable insights into the distribution of project texts and project measurements while offering recommendations for proposed projects. Something to keep in mind when using our work is that we assume the “areas of interest” already exist in the data. As the mining phase of our process does not make any predictions for “areas of interest,” there will be no recommendations of places not already measured.

References:

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003.

[2] Ricardo J. G. B. Campello, Davoud Moulavi, Joerg Sander, “Density-Based Clustering Based on Hierarchical Density Estimates,”*Advances in Knowledge Discovery and Data Mining*, vol. 7819, no. null, pp. 160–172, 2013.

[3] K. Singh, “How to Improve Class Imbalance using Class Weights in Machine Learning?,” *Analytics Vidhya*, Jul. 06, 2023. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

UNIVERSITY of VIRGINIA | DATA SCIENCE