COURSEWORK SUMMARY FOR AM41UD

Student Name: OKUNNUGA ISRAEL OLAMIDE

Student ID: 230427033

Programme: MSc Data Science

1. INTRODUCTION

The identification of B-cell epitopes is a crucial process for several medical and immunological processes including vaccine development, disease diagnosis and prevention (Ashford *et al.*, 2021). In recent years, Machine learning methods have been adopted as the main method for epitope prediction than previously used propensity scales due to better prediction performance (Ashford *et al.*, 2021). This report is a summary of the development of an efficient data mining pipeline to predict Linear B-cell epitopes for the protozoan parasite, Trypanosoma cruzi, the causative agent of Chagas disease. It is a methodical pipeline built from Exploratory Data Analysis (EDA) and Data Pre-Processing (DPP), modelling and hyperparameter tunings, eventually resulting in a final ensemble optimized for balanced accuracy. The goals were to:

- Conduct a reasonable exploration of the data to understand its attributes (features distributions, scales, and dependencies) with adequate visualization.
- Take different pre-processing decisions based on the insights derived from the EDA, including possible feature reduction strategies suited to high dimensional embeddings.
- Treat any extreme class imbalance through sampling and ensemble methods, and
- Build, tune, and compare different classification models, whilst addressing the extreme class imbalance (0.7% positive), before selecting a final pipeline based on nested cross-validation (CV) balanced accuracy.

2. EDA and DATA PRE-PROCESSING (DPP)

Data Overview and Quality Checks: After loading the provided csv file (df.csv), investigation of the feature types showed it contained Info_ and feat_ columns. The 'Info_' columns were irrelevant for modelling, but the 'Info_group' was to be used as the grouping variable for splitting. The 'feat_' columns were continuous floats, and they were to be used for modeling. No missing values, duplicates, or placeholder values were discovered. The

target distribution showed 44,718 negatives (-1) and 332 positives (1), revealing severe class imbalance (0.7%).

Exploratory Visualization: Univariate histograms that were used for representative features (e.g. feat_entropy, feat_AAtypes_Small) exposed varying skewness, indicating that Tree-based models would be best suited for this task. A correlation heatmap of the first 50 features revealed clusters of highly correlated embeddings indicating that dimensionality reduction would be necessary.

Group-Aware Split, Scaling, and Outlier Detection: Using Info_group as the grouping variable, GroupShuffleSplit was utilized to split the data to avoid data leakage. Thousands of extreme values were flagged when Z-score and IQR methods were applied to detect outliers, necessitating the choice of RobustScaler-which uses median centering and IQR scaling-to scale the data. Post-scaling checks confirmed medians to be approximately 0 and 1, meaning each feature was now centered on its median and directly comparable in magnitude.

3. FEATURE REDUCTION

Due to the high dimensionality and multicollinearity of the data, PCA was employed to cut the features whilst retaining components for 95% and 90% explained variance. Empirical testing conducted on PCA-reduced data using baseline classifiers performed poorly, mostly due to class imbalance as the classes weren't linearly separable in the feature subspace that PCA captured. A scatterplot used for visualization confirmed that the minority class (-1) was deeply embedded in the cloud of the majority class (1), indicating that nonlinear model ensemble methods would be necessary to capture minority class signals.

4. MODELLING AND ASSESSMENT

Comparison of Random Forest classifiers (with and without SMOTE and class weights) with and without PCA informed the need to drop PCA and also to not explore further on dimensionality reduction. XGBoost was chosen as a suitable classifier because of its awareness of severe imbalances in datasets. XGB + SMOTE pipelines with thresholds of 0.01 and 0.10 gave balanced accuracies of approximately 0.57 and 0.60, establishing a strong baseline. Using GroupKFold and per-fold threshold optimization, evaluation and

comparisons of the performances of SMOTE, ADASYN, SMOTE-Tomek, SMOTEENN, Cluster centroids, and Balanced Bagging was carried out. Their balanced accuracies clustered around 0.52-0.59 with the exception of Balanced Bagging that was 0.77, making it the best choice for selection, especially with its internal balanced sampling ability. After achieving an average best threshold of 0.10 for Balanced Bagging, an ensemble of XGBoost estimators was applied, under-sampling within each bag. Extensive and robust tuning created a final pipeline with refined hyperparameters: estimator_max_depth= 6, estimator_learning_rate= 0.12, estimator_n_estimators= 150 (150 XGB trees), n_estimators = 25 bags, and sampler = None. This fully tuned pipeline gave a Cross-Validation balanced accuracy of approximately 0.68, and the Outer GroupKFold (5 folds) over the final pipeline (scaling and bagging) provided a robust generalization estimate of 66.1% +- 3.9% balanced accuracy.

5. CONCLUSIONS AND DISCUSSION

This study:

- Showed that group-aware splitting is important to prevent dilated metrics when peptides have a common biological context.
- Revealed that ensemble methods are useful for uncovering subtle epitope signals in high-dimensional protein embeddings.
- Highlighted that Balanced Bagging of XG boost classifiers remarkably outperforms single classifiers or standalone sampling when dealing with extremely imbalanced data, boosting balanced accuracy by over 20% over baseline.
- Lends credence to the point that unsupervised learning enables state-of-the-art supervised predictions of protein sequences and improves state-of-the-art features for long range contact prediction (Rives *et al.*, 2021).

REFERENCES

1. Ashford, J. *et al.* (2021) 'Organism-specific training improves performance of linear B-cell epitope prediction,' *Bioinformatics*, 37(24), pp. 4826–4834. https://doi.org/10.1093/bioinformatics/btab536.
2. Rives, A. *et al.* (2021) 'Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,' *Proceedings of the National Academy of Sciences*, 118(15). https://doi.org/10.1073/pnas.2016239118.