

<p>Synthetic Methodology: SMOTENC</p> <p>script: BinarySMOTENC.py</p> <p>description: Generated synthetic samples for minority class (income \geq \$50K) using SMOTENC algorithm to address class imbalance in the UCI Adults dataset.</p> <p>parameters: $k_neighbours \rightarrow 5$, $ratio_limit \rightarrow 0.8$, $random_state \rightarrow 43$, $categorical_features \rightarrow [workclass, education, marital_status, occupation, relationship, age, race]$</p> <p>Privacy Evaluation: Nearest Neighbor Distance Analysis</p> <p>script: synthetic_data_validation.py</p> <p>privacy_metrics: $avg_nn_distance \rightarrow 0.208$, $min_nn_distance \rightarrow 0.102$, $privacy_threshold \rightarrow 0.1$, $duplicate_count \rightarrow 0$, $privacy_status \rightarrow PASS$</p> <p>Fidelity Evaluation: Distribution Similarity Analysis</p> <p>script: BinaryDatasetValidation1.py</p> <p>fidelity_metrics: $kl_divergence_max \rightarrow 0.0181$, $js_divergence_max \rightarrow 0.0039$, $tvd_max \rightarrow 0.0405$, $coverage_metric \rightarrow 82.36\%$, $avg_emd \rightarrow 0.196$, $r4_compliance \rightarrow PASS$</p> <p>Bias Evaluation: Subgroup Distribution Analysis</p> <p>script: fairness_evaluation.py</p> <p>relevant_groups: $\{race, sex, native_country\}$</p> <p>bias_metrics: $tvd_race \rightarrow 0.0405$, $tvd_sex \rightarrow 0.0116$, $tvd_native_country \rightarrow 0.0374$, $max_allowed_tvd \rightarrow 0.1$, $fairness_status \rightarrow PASS$</p> <p>General Info</p> <p>timestamp: 2025-04-04T15:30:00Z</p> <p>creator: John Smith</p> <p>original dataset: uci_adults_income_train</p> <p>output datasets: $\langle original_train, augmented_train, test_set \rangle$</p>
--

Figure 1: Synthetic Data Model Card