**Checklist of recommendations for trustworthy AI according to requirement and lifecycle (P: Data preparation, D: Model development, U: Deployment & Use, M: Management, X: Not part of the life cycle)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CHECKLIST OF TRUSTWORTHY AI RECOMMENDATIONS *based on EU Trustworthy AI guidelines and this work’s framework*** | | | | |
| **Stage** | **Requirement** | **Item** | **Checklist item** | **Satisfied** |
|  | **Data collection and metadata** | | | |
| X | Data types and ranges | M1a | Data types have been identified |  |
| M1b | The valid ranges or domain for each variable are defined |  |
| Data provenance | M2 | Metadata includes information about the origin of data |  |
| Target variable | M3a | The target variable is defined |  |
| M3b | The positive class is defined |  |
| Sensitive Variables | M4a | The possible sensitive variables have been identified and provided |  |
| M4b | A sensitive variable has been selected to balance its instances |  |
| Identification Variables | M4 | The identification variables are defined |  |
|  | **Privacy and data governance** | | | |
| P | Anonymization | G1 | The data does not contain any identifying information |  |
| Standardisation | G2 | The information is adapted to the standards of the context |  |
| Data quality control | G3a | A data quality assessment has been carried out based on well-defined dimensions |  |
| G3b | Data variability across data sources and over time has been evaluated |  |
| M | Changelogs | G4 | A log system changes exists |  |
|  | **Diversity, non-discrimination and fairness** | | | |
| P | Sensitive exploratory analysis | F1 | A sensitive exploratory analysis has been done to find possible relationships associated with sensitive variables in the dataset |  |
| Bias mitigation | F2a | There is implemented a bias mitigation pre-processing method |  |
| D | F2b | There is implemented a bias mitigation in-processing method |  |
| U | F2c | There is implemented a bias mitigation post-processing method |  |
| D | Sensitive variables performance | F3 | Model performance is evaluated specifically for each sensitive variable |  |
| U | Fairness monitoring | F4 | The system includes a fairness monitoring system in the deployment |  |
|  | **Transparency** | | | |
| P | General exploratory analysis | T1 | A general exploratory analysis has been done to find possible relationships or further problems in the dataset variables |  |
| D | Design description | T2 | System design steps are considerably explained |  |
| Explainability plots | T3a | Explainability techniques or visualizations are included to support the results of the training model. |  |
| U | T3b | Explainability techniques or visualizations are included to support the results of the deployed model |  |
| Disclaimer | T4 | The system includes a disclaimer with its features and limitations |  |
| M | Output Documentation | T5 | Outputs are carefully recorded |  |
| Incident Sharing | T6 | There has been provided a way to report incidents |  |
|  | **Technical robustness and safety** | | | |
| P | Dimensionality | R1 | Potential dimensionality problems are handled with feature selection or dimensionality reduction methods |  |
| Class balancing | R2a | There is implemented a class balancing pre-processing method |  |
| R2b | There is implemented a class balancing in-processing method |  |
| D | Hyperparameter optimization | R3 | Extensive hyperparameter optimization has been performed to optimize model performance |  |
| Well-formed evaluation | R4 | Metrics evaluation provides robust results appropriate to the context |  |
| Attack prevention | R5 | There have been implemented methods to prevent the model for attacks of different nature |  |
| Uncertainty Quantification | R6a | Uncertainty is quantified and reported in metric |  |
| U | R6b | Uncertainty is quantified in predictions |  |
| R6c | Predictions have a threshold of uncertainty above which are not made |  |
| Dataset shift monitoring | R7 | Dataset shifts are considered and handled in the development & monitored in further use |  |
| Retraining | R8 | A retraining system for external implementations has been considered. |  |

*\* Item structure: Requirement – Index – Subsection Index | EXAMPLE: T3a (Requirement: Transparency – Index: 3 – Subsection: a)*

*\*\* This checklist is open to new methods for the current and remaining requirements*

Part of the work **‘A trustworthy health AI development framework with example code pipelines’ by Carlos de-Manuel-Vicente, David Fernández-Narro, Vicent Blanes-Selva, Juan M García-Gómez and Carlos Sáez.** [**https://github.com/bdslab-upv/trustworthy-ai**](https://github.com/bdslab-upv/trustworthy-ai)