Model Card: Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction

Model Details

- **Developer**: Joshna Medisetty , ISE 244 Project, Spring 2025
- **Model Date**: April 2025
- **Model Version**: 1.0
- **Model Type**: Comparative supervised classification (tabular data)
- **Algorithms**:
    a. Logistic Regression
    b. Random Forest
    c. Extra Trees Classifier
    d. Support Vector Machine (RBF kernel)
    e. XGBoost
    f. Deep Neural Network (Multi-layer Perceptron)
- **Frameworks**: scikit-learn, XGBoost, TensorFlow/PyTorch
- **Contact**: joshna.medisetty@sjsu.edu
- **License**: For educational use only

Intended Use

- **Primary Intended Uses**:
    - Research and educational demonstration of ML model comparison for binary disease prediction.
    - Preliminary risk screening for diabetes in population health studies.
- **Primary Intended Users**: Data science students, ML researchers, public health analysts.
- **Out-of-scope Uses**:
    - Not for direct clinical decision-making or diagnosis.
    - Not validated for populations outside the Pima Indian cohort or for use with non-tabular data.

Factors

- **Relevant Factors**:
    - Age group, BMI category, gender (if available), and other demographic/phenotypic subgroups.
    - Data quality (missing values, outliers).
- **Evaluation Factors**: Model performance is reported overall and, where possible, disaggregated by age and BMI subgroups to highlight potential disparities

Metrics

- **Performance Metrics**:
    - Accuracy, Precision, Recall, F1-score, AUC-ROC.
    - Confusion matrix metrics (False Positive Rate, False Negative Rate) for fairness analysis
- **Decision Threshold**: Default threshold at 0.5 for all probabilistic models.
- **Reporting**:
    - All metrics reported on a held-out test set (30% of data), with 10-fold cross-validation on the training set.
    - 95% confidence intervals via bootstrapping.

Evaluation Data

- **Dataset**:
    - Pima Indians Diabetes Dataset (UCI ML Repository)
    - 768 samples, 8 features, binary outcome (diabetes yes/no)
- **Motivation**:
    - Widely used benchmark for binary disease prediction ([Liao et al., 2021, Sec. 2]).
- **Preprocessing**:
    - Median imputation for missing values, standard scaling, SMOTE for class balancing.

Training Data

- **Source**: Same as evaluation data (no additional external data).
- **Demographics**: All female, Pima Indian ancestry, 21 years and older.
- **Distributional Caveats**: Not representative of all ethnicities, genders, or age ranges.

Quantitative Analyses

- **Unitary Results**: All models evaluated on overall test set and on subgroups (e.g., age <30 vs. ≥30, BMI categories).
- **Intersectional Results**: Where sample size allows, performance reported for intersections (e.g., older/obese subgroup).
- **Findings**:
    - XGBoost and DNN generally outperform simpler models in AUC-ROC and F1, but at higher computational cost.
    - Logistic Regression provides competitive baseline with greater interpretability.

Ethical Considerations

- **Sensitive Data**: Dataset contains health and demographic data; privacy respected by using only public, de-identified data.
- **Human Impact**: Model is not intended for clinical use; risks include misclassification leading to false reassurance or unnecessary anxiety.
- **Bias and Fairness**: Potential for bias due to limited population diversity and class imbalance.
- **Mitigations**: SMOTE used for balancing; subgroup performance reported to surface disparities.
- **Harms**: Risk of overfitting to benchmark dataset, poor generalization to other populations.

Caveats and Recommendations

- **Internal Validity**:
    - All models compared using the same splits and preprocessing; hyperparameters tuned via grid search.
    - Care taken to avoid test set leakage and overfitting.
- **External Validity**: Results may not generalize to other datasets, populations, or real-world clinical settings.
- **Further Testing**:
    - Recommend evaluation on more diverse datasets and with additional demographic/phenotypic subgroups.

- Future work should address interpretability (e.g., SHAP/LIME) and calibration.

References

- Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model Cards for Model Reporting.
- Raji, I.D., et al. (2022). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.
- Liao, T.I., et al. (2021). Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning.