تاریخ برگزاری همایش
۱۴۰۴/۰۲/۲۴

رویداد ملی
کاربرد هوش مصنوعی در عصر نوین
با محوریت: امنیت ملی، صنایع و معادن، محیط زیست، حقوق، شهرسازی

دانشگاه آزاد اسلامی
سیرجان

DIABETES PREDICTION WITH MACHINE LEARNING: A GENETIC ALGORITHM APPROACH

Ali Mirzaei, Dr. Jalal Nasiri
Department of Computer Science
Ferdowsi University of Mashhad
Mashhad, Iran
alipmirzaei@gmail.com

Abstract

This paper presents a comprehensive study on predicting diabetes using machine learning techniques. We explore a range of algorithms, including SVM, TwinSVM, Decision Tree, Random Forest, Bagging with SVM, XGBoost, and a Decision Tree model enhanced with Genetic Algorithm (GA) based feature selection, to identify the most effective model for this classification task. The study emphasizes the importance of feature selection, employing a Genetic Algorithm (GA) to determine the optimal subset of features. Data preprocessing steps, such as handling class imbalance and feature scaling, are also detailed. Performance evaluation is conducted using accuracy, precision, recall, and F1-score. The results demonstrate that the XGBoost and Decision Tree with GA achieves the highest accuracy.

Index Terms–diabetes prediction, machine learning, genetic algorithm, feature selection

1 INTRODUCTION

Diabetes is a chronic disease that affects millions of people worldwide [15]. Early and accurate prediction of diabetes can significantly improve patient outcomes by enabling timely interventions and lifestyle modifications [16]. Machine learning (ML) techniques have shown great promise in developing predictive models for various medical conditions, including diabetes [3, 12].

This paper investigates the application of several ML algorithms to predict diabetes based on the ”Diabetes Prediction Dataset” [15]. We focus on the following key aspects:

1. **Data Preprocessing:** We describe the steps taken to prepare the dataset for model training, including handling categorical features, addressing class imbalance, and scaling numerical features.

2. **Feature Selection:** A Genetic Algorithm (GA) [5, 6] is implemented to identify the most relevant features for diabetes prediction, potentially improving model performance and reducing complexity.

3. **Model Training and Evaluation:** We train and evaluate six different ML models: SVM, TwinSVM, Decision Tree, Random Forest, Bagging with SVM, and XGBoost. In addition, we analyze a Decision Tree model enhanced with GA-based feature selection.

4. **Performance Comparison:** We compare the performance of the models using metrics such as accuracy, precision, recall, and F1-score.

2 DATASET AND PREPROCESSING

The study utilizes the ”Diabetes Prediction Dataset” [15], which contains various patient attributes, including age, gender, BMI, smoking history, HbA1c level, blood glucose level, hypertension, and heart disease. The target variable is 'diabetes,' indicating whether a patient has diabetes (1) or not (0).

2.1 Data Preprocessing

The following preprocessing steps were performed:

1. **Encoding Categorical Features:** The 'gender' column was label-encoded into a binary 'is_female' column. The 'smoking_history' column was one-hot encoded, and the 'No Info' category was dropped to avoid redundancy.

2. **Handling Class Imbalance:** The dataset exhibited class imbalance, with significantly more instances of non-diabetic patients. SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors) was applied [16, 17] to balance the class distribution in the training set.

3. **Data Sampling:** A random sample of 10,000 instances was selected from the resampled training data for model training.

4. **Feature Scaling:** The 'StandardScaler' from scikit-learn was used to standardize the numerical features, ensuring that they have zero mean and unit variance. This step is crucial for algorithms sensitive to feature scales, such as SVM.

2.2 Data Exploration

as shown in **figures 1 and 2** the distribution of each feature, separated by diabetes status (whether the individual has diabetes or not).

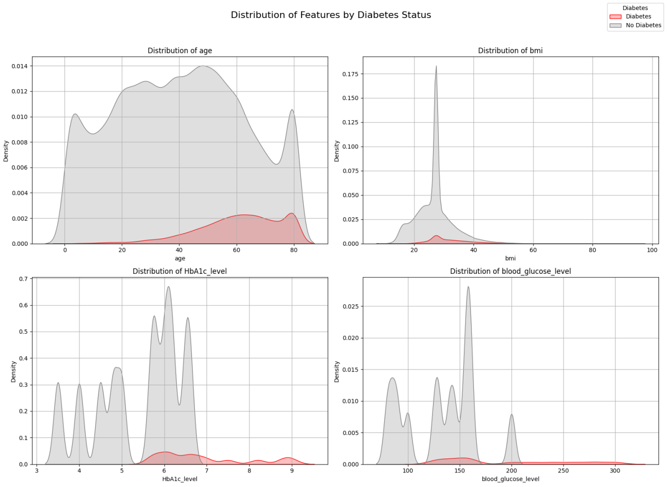

Figure 1: Distribution of Numerical Features

1. **Numerical Features:**

   - **Age:** The distribution for individuals with diabetes is skewed towards older ages, suggesting that older individuals tend to have a higher risk of diabetes.

   - **BMI:** The BMI distribution for the diabetic group is shifted to the right, indicating that higher BMI is associated with an increased risk of diabetes. There are also some noticeable outliers in the non-diabetic group with very high BMI.

   - **HbA1c_level:** Individuals with diabetes have significantly higher HbA1c levels, with a clear separation between the two groups' distributions. This suggests that HbA1c level is a strong predictor of diabetes.

   - **Blood_glucose_level:** diabetic individuals similar to HbA1c level exhibit notably higher blood glucose levels compared to non-diabetic individuals.

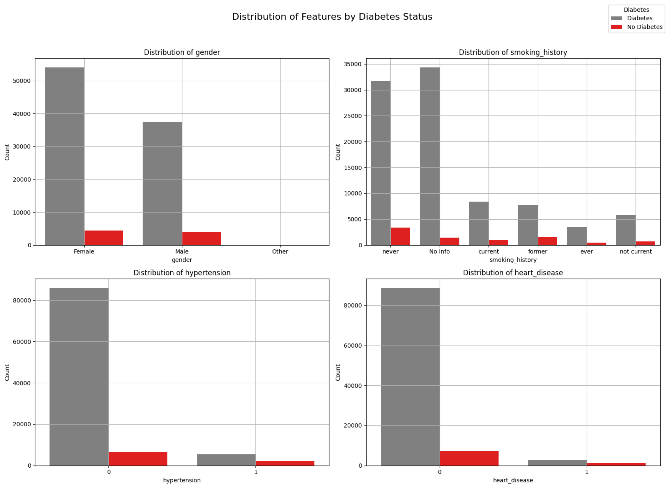

Figure 2: Distribution of Categorical Features

2. **Categorical Features:**

   - **Gender:** The distribution of diabetes appears relatively similar between males and females, although there is a slightly higher proportion of females in the dataset.
   - **Smoking_history:** The 'No Info' category constitutes a large portion of the data, making it difficult to draw definitive conclusions about the correlation between smoking history and diabetes. Further investigation or data imputation might be necessary.
   - **Hypertension:** A higher proportion of individuals with hypertension have diabetes compared to those without hypertension.
   - **Heart_disease:** Individuals with heart disease also show a higher prevalence of diabetes.

Figure 3 illustrates the correlation matrix of the features in the resampled training dataset. The heatmap visually represents the pairwise correlations between different features, with darker colors indicating stronger correlations. Notably, strong positive correlations are observed between the selected features (HbA1c_level and blood_glucose_level) and the target variable (diabetes), justifying their selection by the Genetic Algorithm.

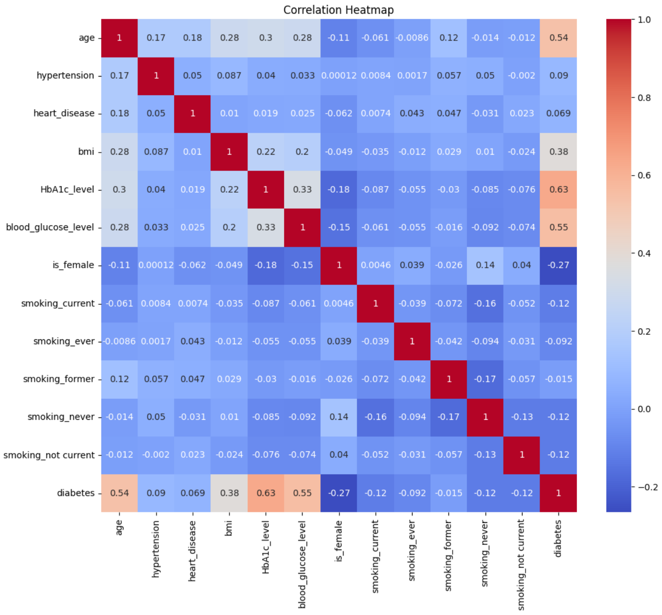

Figure 3: Correlation Heatmap of Features in the Resampled Training Dataset

3 METHODOLOGY

3.1 Machine Learning Models

This study evaluates the performance of the following machine learning models:

1. **SVM (scikit-learn):** An ‘SVC‘ from scikit-learn was trained, and its hyperparameters were tuned using ‘GridSearchCV‘ [19] with 5-fold cross-validation and then The best model was selected based on accuracy.
2. **TwinSVM:** A custom ‘TwinSVM‘ class was implemented based on the Twin Support Vector Machine algorithm. Hyperparameter tuning was performed manually using nested loops.
3. **Decision Tree (DT):** A ‘DecisionTreeClassifier‘ was trained, and its hyperparameters were tuned using ‘GridSearchCV‘ [19] with 5-fold cross-validation and The best model based on accuracy was selected.
4. **Random Forest (RF):** A ‘RandomForestClassifier‘ was trained, and its hyperparameters were tuned us-

ing ‘GridSearchCV‘ [19] with 5-fold cross-validation. The best model based on accuracy was selected.

5. **Bagging with SVM:** A ‘BaggingClassifier‘ was used with an ‘SVC‘ (Support Vector Classifier) as the base estimator. The ‘SVC‘ used a linear kernel, C=10, and gamma=’scale’.

6. **XGBoost:** An ‘XGBClassifier‘ was trained with default parameters.

3.2 Genetic Algorithm for Feature Selection

A Genetic Algorithm (GA) [5, 6] was employed to select an optimal subset of features for diabetes prediction. The GA was implemented using the ‘deap‘ library in Python [18]. The following components were defined:

- **Individual:** A binary list representing a feature subset, where 1 indicates the feature is selected, and 0 indicates it is not.

- **Fitness Function:** A ‘DecisionTreeClassifier‘ was trained on the selected features, and the accuracy on the test set was used as the fitness score.

- **Population:** A population of 20 individuals was initialized.

- **Genetic Operators:**
  - **Selection:** Tournament selection with a tournament size of 3.
  - **Crossover:** Two-point crossover with a probability of 0.7.
  - **Mutation:** Bit-flip mutation with an independent probability of 0.2.

The GA was executed for 10 generations, and the individual with the highest fitness score in the final population determined the selected features.

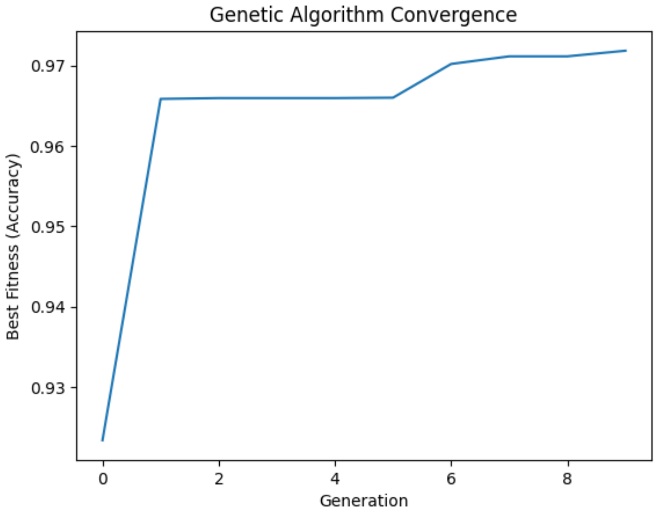

Figure 4: Genetic Algorithm Convergence

Figure 4 depicts the convergence of the Genetic Algorithm over 10 generations. The plot shows the best fitness score (accuracy) achieved in each generation. We can observe that the GA quickly converges to a high accuracy level, indicating the effectiveness of the feature selection process.

The Genetic Algorithm selected the following features as the most relevant for diabetes prediction:

- HbA1c_level
- blood_glucose_level

3.3 Decision Tree with GA-Selected Features

The best-performing Decision Tree model from the hyperparameter tuning step was retrained using only the features selected by the Genetic Algorithm to assess the impact of feature selection on model performance.

4 RESULTS AND DISCUSSION

The performance of each model was evaluated on the test set using accuracy, precision, recall, and F1-score. The results are summarized in Table I.

Table I: Performance of Different Models

Model	Accuracy	Precision	Recall	F1-Score
SVM (scikit-learn)	0.8709	0.93	0.87	0.89
TwinSVM	0.8761	0.93	0.88	0.90
Decision Tree	0.9141	0.94	0.91	0.92
Random Forest	0.9018	0.94	0.90	0.91
Bagging with SVM	0.8763	0.93	0.88	0.90
XGBoost	0.9720	0.97	0.97	0.97
GA + Decision Tree	0.9719	0.97	0.97	0.97

4.1 Model Performance Comparison

The models are presented in the order they were implemented in this study, starting with SVM-based approaches, followed by tree-based methods, and concluding with the GA-enhanced Decision Tree.

SVM (scikit-learn): Achieved an accuracy of 0.8709, with high precision but slightly lower recall, indicating a good balance between correctly identifying positive cases and minimizing false positives.

TwinSVM: Showed a marginal improvement in accuracy (0.8761) and similar precision and recall to the standard SVM.

Decision Tree: Outperformed both SVM models with an accuracy of 0.9141, and high precision (0.94), recall (0.91), and F1-score (0.92), demonstrating its effectiveness in capturing the underlying patterns in the data.

Random Forest: Performed comparably to the Decision Tree, with a slightly lower accuracy of 0.9018, indicating robust performance across different tree-based models.

Bagging with SVM: Achieved an accuracy of 0.8763, similar to the TwinSVM, but did not surpass the performance of the tree-based models.

XGBoost: Demonstrated the highest accuracy among all models at 0.9720, with excellent precision, recall, and F1-score (all 0.97), underscoring the power of gradient boosting algorithms in this predictive task.

GA + Decision Tree: This model, which utilized only the features selected by the Genetic Algorithm, achieved an accuracy very close to XGBoost (0.9719), with equally high precision, recall, and F1-scores. This highlights the effectiveness of combining GA-based feature selection with a DT classifier.

5 CONCLUSION

This study demonstrated the effectiveness of various machine learning algorithms in predicting diabetes based on patient attributes. Notably, the XGBoost model and the Decision Tree trained with GA-selected features achieved the highest accuracies, both approximately 0.972. The Genetic Algorithm's ability to identify a highly discriminative subset of features (HbA1c_level and blood_glucose_level) proved beneficial, enhancing the Decision Tree's performance to match that of the more complex XGBoost model. The strong performance of these models, particularly in terms of precision and recall, suggests their potential for use in real-world clinical settings to aid in early diabetes diagnosis [15]. However, further validation on larger and more diverse datasets is necessary to confirm these findings.

Future work could explore more advanced feature engineering techniques, investigate a wider range of hyperparameters for each model, and incorporate additional data sources to further enhance the accuracy and robustness of the predictive models. Additionally, exploring other ensemble methods like stacking or voting classifiers, as well as deep learning models, might lead to further performance improvements.

6 ACKNOWLEDGMENT

This work was completed as part of the Computational Intelligence course at Ferdowsi University of Mashhad during the Fall 2024 semester. I would like to express my gratitude to the teaching assistants of the course for their support and guidance throughout the project. I also want to thank the creators of the Diabetes Prediction Dataset for making this valuable resource available for research.

7 REFERENCES

References

[1] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[2] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2001.

[4] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[5] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA: MIT Press, 1992.

[6] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press, 1998.

[7] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT 2010*, 2010, pp. 177–186.

[8] J. D. Schaffer, R. A. Caruana, L. J. Eshelman, and R. Das, "A study of control parameters affecting online performance of genetic algorithms for function optimization," in *Proc. 3rd Int. Conf. Genet. Algorithms*, 1989, pp. 51–60.

[9] G. E. Hinton, "Deep learning," *Scholarpedia*, vol. 2, no. 8, p. 2949, 2007.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[12] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, 2007.

[13] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco: Morgan Kaufmann, 1991.

[14] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Mach. Learn.*, vol. 3, no. 2, pp. 95–99, 1988.

[15] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://archive.ics.uci.edu/ml

[16] G. M. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes and unequal error costs?" in *Proc. DMIN*, 2007, pp. 35–41.

[17] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[18] F.-A. Fortin et al., "DEAP: Evolutionary Algorithms Made Easy," *J. Mach. Learn. Res.*, vol. 13, pp. 2171–2175, 2012.

[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.