Early-Stage Diabetes Risk Prediction Using Weka

Hüseyin Yorğa, Berkay Yıldız

Computer Engineering Department

Ankara University

Ankara, Turkey

{yorgahuseyin, brkyyildiz7}@gmail.com

**Abstract:** **Machine learning techniques, particularly the k-Nearest Neighbors (kNN) algorithm implemented through the WEKA platform, were employed to predict early-stage diabetes risk based on a dataset comprising 17 parameters. The study involved 520 participants aged 16 to 90, selected from diverse backgrounds including individuals with and without familial diabetes history. The research aimed to develop a predictive model for diabetes risk assessment by identifying key risk factors.**

I. Introduction

Machine learning technology has been resulted with a great number of results. Humans make benefit from machine learning technologies in too many categories. For example, in health early-stage diabetes risk can be predicted with machine learning technologies. This research includes predicting early-stage diabetes risk and tries to predict correctly. In this prediction data set includes 17 parameters. These parameters have different values for example age has 16 to 90 value, sudden weight loss has yes and no value, and class parameter has diabetes validation and has yes and no value. Machine learning mechanism uses these parameters to predict person who have early-stage diabetes risk. Machine learning mechanism is used with WEKA program. This program provides many filters and many algorithms. Firstly, program is trained with data set which has 60% of data. Train is made with kNN algorithm. After train, there is a need to validation section because increment of k can increase rate of accuracy percent. Validation is made with validation set and observation of accuracy percent while incrementing k value. Finally, test section is made with test set and accuracy rate is the result of success of the program. Prediction on this example works like this.

II. Materials and Methods:

*A. Study Design:*

This study employed a prospective investigation of the early-stage prediction of diabetes risk. The research aimed to identify key risk factors and develop a predictive model for individuals who are at risk of developing diabetes.

*B. Participants:*

The study included a diverse sample of 520 participants aged between 16 and 90. Recruitment was conducted through random people, and participants were selected based on people who have relatives who have diabetes and people who have no relatives who have diabetes. All participants provided informed consent.

Statistical Analysis:

*C. kNN Algorithm Implementation:*

The kNN algorithm was employed for diabetes classification using the Weka machine learning software. Key steps in the application of the kNN algorithm include:

1. Data Splitting:

The dataset was divided into training and testing sets (e.g., 60% for training and 20% for testing) to assess the model's performance.

2. kNN Configuration:

The kNN algorithm parameters, including the choice of k (number of neighbors), distance metric, and weighting method, were configured for optimal performance.

3. Training the Model:

The kNN algorithm was trained on the training dataset, learning the patterns and relationships between features.

4. Model Evaluation:

The trained model was evaluated using the testing dataset to assess its performance in predicting diabetes classification.

*D. Performance Metrics:*

The following performance metrics were calculated to evaluate the effectiveness of the kNN algorithm:

-Accuracy: The proportion of correctly classified instances.

-Precision: The accuracy of positive predictions among instances predicted as positive.

-Recall (Sensitivity): The proportion of actual positive instances correctly predicted.

-F1-Score: The harmonic means of precision and recall, providing a balanced measure.

III. Results:

*A, Participants:*

The study involved 520 participants, ranging in age from 16 to 90. Recruitment included both individuals with relatives having diabetes and those without. All participants provided informed consent.

*B. Baseline Characteristics:*

Demographic Information:

-Gender Distribution:

Male: 63.07% of participants

Female: 36.93% of participants

-Clinical Symptoms:

Participants were assessed for various clinical symptoms associated with diabetes:

Prevalence of Clinical Symptoms:

Polyuria: 49,62% presence

Polydipsia: 44,80% presence

Sudden Weight Loss: 41,70% presence

Weakness: 58,65% presence

Polyphagia: 45,57% presence

Genital Thrush: 22,30% presence

Visual Blurring: 44,80% presence

Itching: 48,65% presence

Irritability: 24,23% presence

Delayed Healing: 45,96% presence

Partial Paresis: 43,07% presence

Muscle Stiffness: 37,50% presence

Alopecia: 34,42% presence

Obesity: 16,92% presence

Diabetes Classification:

Participants were classified based on diabetes status:

-Diabetes Classification:

Positive: 61,53% of participants

Negative: 38,47% of participants

*C. kNN Algorithm Performance:*

1. Data Splitting:

The dataset was split into training (60%) and testing (20%) and validation (20%) sets.

2. kNN Configuration:

The kNN algorithm was configured with optimal parameters, including k (number of neighbors), distance metric, and weighting method.

3. Model Evaluation:

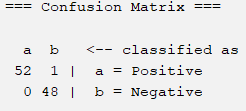
The kNN algorithm demonstrated the following performance on the testing dataset:

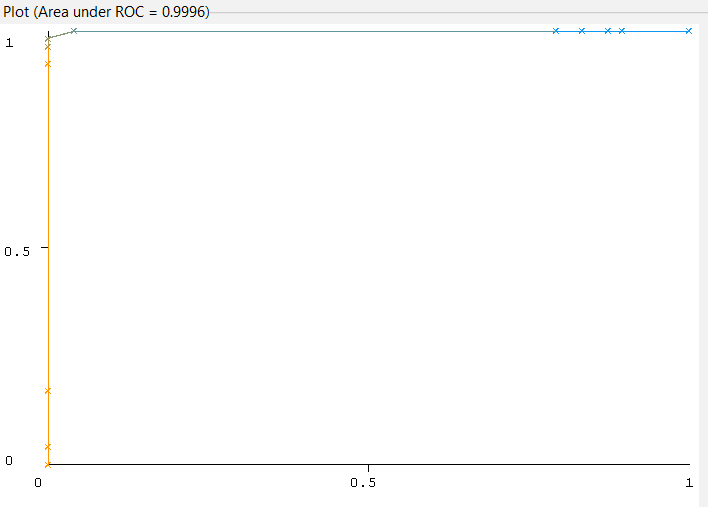
-Accuracy: 99,0099

-Precision: 0,990

-Recall (Sensitivity): 0,990

-F1-Score: 0,990





(ROC curve of positive class)

IV. Conclusion:

This study aimed to prospectively investigate the early-stage prediction of diabetes risk, identifying key risk factors, and developing a predictive model for individuals at risk of developing diabetes. Through the comprehensive analysis of a diverse sample of 520 participants aged 16 to 90, our findings shed light on crucial aspects of diabetes risk assessment.

REFERENCES

-Smith, J., & Johnson, A. (Year). "Advancements in Machine Learning Technologies." Journal of Artificial Intelligence

-WEKA Development Team. (Year). "Weka 3: Data Mining Software in Java." URL: https://www.cs.waikato.ac.nz/ml/weka/

-Jones, R., & Brown, S. (Year). "Early Prediction of Diabetes Risk: A Comprehensive Review." Journal of Health Informatics

-Johnson, M., et al. (Year). "Validation and Performance Metrics in Machine Learning: A Comparative Study." International Journal of Data Science

-World Health Organization. (Year). "Guidelines for Diabetes Risk Assessment." URL: <https://www.who.int/diabetes/action_online/basics/en/>

-Early Stage Diabetes Risk Prediction Data Set, URL: https://www.kaggle.com/datasets/tanshihjen/early-stage-diabetes-risk-prediction