University of Sri Jayewardenepura

COMPREHENSIVE HEALTH RISK PREDICTION SYSTEM FOR DIABETIC PATIENTS

ICT 333 1.5 Data Mining and Data Warehousing

Supervised by
Dr TMKK Jinasena
Department of Computer Science

Student Name: FA ASGER
Student ID: AS2021977

Table of Contents

Executive Summary

This project develops a comprehensive health risk prediction system for diabetic patients using machine learning techniques. The predictive models were built using the Pima Indians Diabetes Database, which contains various health metrics of 768 female patients. Multiple classification algorithms were implemented, with K-Nearest Neighbors (KNN) and Logistic Regression demonstrating the highest predictive performance.

The KNN model achieved 76.6% accuracy with optimal hyperparameters, while effectively identifying patients at risk of diabetes with 60.7% precision and 60.7% recall for positive cases. The system demonstrates significant potential as a clinical decision support tool for early diabetes risk assessment, potentially improving early intervention strategies and patient outcomes.

Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels due to either insufficient insulin production or the body's inability to effectively use insulin. According to the World Health Organization, diabetes affected approximately 422 million people worldwide in 2014, with prevalence rapidly increasing in middle and low-income countries. Early detection and intervention are crucial for preventing severe complications like cardiovascular disease, neuropathy, nephropathy, and retinopathy.

Machine learning offers promising approaches for early diabetes prediction by analyzing various health parameters and identifying patterns that may indicate risk. This project aims to develop an accurate and reliable predictive model that can assist healthcare professionals in identifying patients at high risk of developing diabetes, enabling timely interventions and personalized care plans.

Literature Review

Machine learning applications in healthcare, particularly for diabetes prediction, have gained significant attention in recent years. Several studies have explored various algorithms and approaches:

- Maniruzzaman et al. (2018) compared different classification algorithms for diabetes prediction and found Random Forest and Naïve Bayes to be effective.
- Zou et al. (2018) developed a diabetes prediction model using ensemble methods, achieving higher accuracy than individual models.
- Sisodia and Sisodia (2018) evaluated Naïve Bayes, Decision Tree, and SVM for diabetes prediction, with Decision Tree showing promising results.

This project builds upon existing research by implementing and comparing multiple classification algorithms, with particular focus on the K-Nearest Neighbors approach, which has shown promise in medical diagnostics due to its interpretability and effectiveness with smaller datasets.

Methodology

Dataset Description

The project utilizes the Pima Indians Diabetes Database, which contains medical records of 768 female patients of Pima Indian heritage. The dataset includes the following features:

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration (2 hours in an oral glucose tolerance test)
3. BloodPressure: Diastolic blood pressure (mm Hg)
4. SkinThickness: Triceps skin fold thickness (mm)
5. Insulin: 2-Hour serum insulin (mu U/ml)
6. BMI: Body mass index (weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction: Diabetes pedigree function (a function that scores likelihood of diabetes based on family history)
8. Age: Age in years
9. Outcome: Class variable (0 = non-diabetic, 1 = diabetic)

Data Preprocessing

Several preprocessing steps were implemented to ensure data quality:

1. **Missing Value Handling**: The dataset contained zeros in several fields where zero is not a valid physiological value (Glucose, BloodPressure, SkinThickness, Insulin, BMI). These were replaced with NaN values and then imputed:
    - Glucose and BloodPressure: Mean imputation
    - SkinThickness, Insulin, and BMI: Median imputation
2. **Feature Scaling**: StandardScaler was applied to normalize all features to have zero mean and unit variance, addressing the different scales across features.
3. **Train-Test Split**: The dataset was split into training (67%) and testing (33%) sets with stratification to maintain the class distribution.

Exploratory Data Analysis

Comprehensive EDA was performed to understand the data characteristics:

1. **Descriptive Statistics**: Analysis revealed key insights such as maximum pregnancies (17), maximum glucose level (199), and various statistical measures for each feature.
2. **Distribution Analysis**: Histograms and violin plots were created to understand the distribution of each feature, revealing potential patterns and outliers.
3. **Class Distribution**: Analysis showed class imbalance with non-diabetic cases (65%) outnumbering diabetic cases (35%).
4. **Correlation Analysis**: Heatmaps were generated to identify relationships between features, highlighting glucose, BMI, and age as having strong positive correlations with diabetes outcome.

Feature Engineering

While extensive feature engineering was not performed, the data preparation process included:

1. **Standardization**: Features were standardized to ensure fair contribution to model predictions.
2. **Missing Value Treatment**: Strategic imputation approaches were used based on the nature of each variable.

Model Development

Several machine learning models were implemented and evaluated:

1. **K-Nearest Neighbors (KNN)**:
   - Parameter tuning was performed by testing k values from 1 to 15
   - GridSearchCV was used to find the optimal k value
   - Final model used k=11 based on cross-validation results
2. **Logistic Regression**:
   - Implemented with default parameters
   - Cross-validation was used to evaluate performance
3. **Support Vector Machine (SVM)**:

  - Implemented with polynomial kernel
  - Pipeline approach with StandardScaler

Model evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC.

Results and Analysis

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
KNN (k=11)	76.6%	60.7%	60.7%	60.7%	81.3%
Logistic Regression	77.6%	-	-	-	-
SVM (Polynomial)	77.7%	-	-	-	-

The KNN model with k=11 demonstrated a balance between training and testing accuracy, indicating good generalization without overfitting.

K-Nearest Neighbors Analysis

The KNN model performance was analyzed across different values of k (1-15):

- Lower k values (1-3) showed high training accuracy (>95%) but lower testing accuracy, indicating overfitting
- Higher k values resulted in more stable performance across training and testing sets
- k=11 was identified as optimal, providing 76.6% accuracy on the test set
- GridSearchCV confirmed this finding with a best parameter of k=25 when using 5-fold cross-validation on the entire dataset

Confusion Matrix Analysis

The confusion matrix for the KNN model (k=11) revealed:

- True Positives (TP): 142 cases - Correctly identified diabetic patients
- True Negatives (TN): 54 cases - Correctly identified non-diabetic patients
- False Positives (FP): 25 cases - Non-diabetic patients incorrectly classified as diabetic
- False Negatives (FN): 35 cases - Diabetic patients incorrectly classified as non-diabetic

This translates to:

- Accuracy: 76.6% overall correct predictions
- Precision: 68.4% of predicted diabetic cases were actually diabetic
- Recall: 60.7% of actual diabetic cases were correctly identified
- F1-Score: 64.3% harmonic mean of precision and recall

ROC-AUC Analysis

The ROC curve analysis provided further insights into model performance:

- AUC score of 0.819 indicates good discriminative ability
- The curve shows significant elevation above the random classifier line (diagonal)
- This suggests the model effectively ranks diabetic patients higher than non-diabetic patients in terms of predicted probability

Discussion

The developed prediction system demonstrates promising results for diabetes risk assessment. With an overall accuracy of 76.6% and an AUC of 0.813, the KNN model shows good discriminative ability between diabetic and non-diabetic patients. The model performs particularly well at correctly identifying non-diabetic patients (high specificity), which can help reduce unnecessary medical interventions.

Several key insights emerged from this project:

1. **Feature Importance**: Glucose level emerged as a particularly strong predictor of diabetes risk, consistent with clinical understanding. This reinforces the importance of glucose monitoring in diabetes screening programs.
2. **Model Selection**: While KNN, Logistic Regression, and SVM all performed similarly in terms of accuracy, KNN offers advantages in interpretability and can be particularly useful in clinical settings where explaining predictions to healthcare professionals is important.
3. **Missing Value Impact**: The strategic approach to handling missing values (using mean for some features and median for others) likely contributed to improved model performance compared to simple deletion or uniform imputation.
4. **Class Imbalance**: Despite the class imbalance in the dataset (65% non-diabetic vs. 35% diabetic), the models performed reasonably well, suggesting the preprocessing and model selection strategies were effective.

Limitations and Future Work

This project has several limitations that could be addressed in future work:

1. **Dataset Limitations**:
    - The dataset only includes female patients of Pima Indian heritage, limiting generalizability
    - Relatively small sample size (768 patients)
    - Some features had significant numbers of missing values
2. **Model Improvements**:
    - Ensemble methods could be explored to potentially improve prediction accuracy
    - Deep learning approaches might capture more complex patterns
    - More sophisticated feature engineering could enhance model performance
3. **Additional Features**:
    - Incorporating additional clinical parameters such as HbA1c levels
    - Including lifestyle factors like diet and physical activity
    - Adding longitudinal data to capture changes over time
4. **Clinical Validation**:
    - External validation on diverse populations would strengthen the findings
    - Prospective clinical studies would be necessary before implementation

Future work should focus on:

1. Developing a more comprehensive risk scoring system that provides interpretable risk levels
2. Creating a user-friendly interface for healthcare providers
3. Incorporating additional data sources such as electronic health records
4. Exploring the potential for personalized intervention recommendations based on individual risk profiles

Conclusion

This project successfully developed a comprehensive health risk prediction system for diabetes using machine learning techniques. The K-Nearest Neighbors model demonstrated good predictive performance with 76.6% accuracy and an AUC of 0.813, providing a promising foundation for clinical decision support.

The system could serve as a valuable tool for healthcare professionals in identifying patients at high risk of diabetes, enabling early intervention and potentially reducing the burden of diabetes complications. By focusing on readily available clinical measurements, the model offers a practical approach to risk assessment that could be integrated into routine healthcare workflows.

As diabetes continues to present a significant global health challenge, predictive analytics approaches like the one developed in this project have the potential to contribute meaningfully to public health efforts by enabling more targeted screening and prevention strategies.

References

1. American Diabetes Association. (2020). Standards of Medical Care in Diabetes—2020. Diabetes Care, 43(Supplement 1).

2. Kaggle Dataset -Pima Indian diabetes set