Diabetes Prediction Based on Support Vector Machine Model

Yundong Xu*, Ying Nie

School of Management, Jiangsu University, Zhenjiang, China, 212013

* Corresponding Author Email: 3210803029@stmail.ujs.edu.cn

Abstract. Diabetes is a chronic condition, the symptoms of which may be relatively mild in the initial stages. Consequently, early prediction is of paramount importance for disease management and treatment. This study aims to employ a Support Vector Machine (SVM) model for the prognostication of diabetes risk. In comparison to conventional prediction techniques, SVMs are adept at capturing complex nonlinear relationships, and possess the capability to handle high-dimensional data and model nonlinearities effectively. By leveraging the strengths of SVMs, it is possible to anticipate the risk of diabetes based on various diagnostic indicators, thereby mitigating the severity of the condition and the risk of complications. The findings of this research offer valuable insights for the management and treatment of diabetes, holding significant application potential for the improvement of diagnosis and therapeutic strategies for diabetic patients.

Keywords: Diabetes, Medical Domain, Support Vector Machine (SVM), Machine Learning, Disease Prediction.

1. Introduction

Diabetes, as a chronic disease, can manifest suddenly, particularly with type 2 diabetes, where the symptoms may be so mild that they go unnoticed for many years. According to a report by the World Health Organization (WHO), the prevalence of diabetes among adults over 18 worldwide reached 8.5% in 2014. By the year 2019, diabetes was directly responsible for 1.5 million deaths, with 48% of these deaths occurring in individuals under the age of 70. Additionally, diabetes-induced kidney diseases accounted for 460,000 deaths, and hyperglycemia was a contributing factor in approximately 20% of cardiovascular disease fatalities. From 2000 to 2019, there was a 3% increase in deaths attributable to diabetes. The death rate from diabetes in low- and middle-income countries rose by 13%. In contrast, the probability of dying between the ages of 30 and 70 from one of the four major noncommunicable diseases—cardiovascular diseases, cancer, chronic respiratory diseases, or diabetes—decreased globally by 22% from 2000 to 2019. Therefore, the early prediction of diabetes risk based on relevant indicators during health check-ups is of utmost importance.

In recent years, data mining and machine learning technologies have shown sustained development and become reliable adjunct tools within the medical field. Data mining methods are widely applied for the preprocessing of medical data and the selection of relevant features, enhancing data quality and the extraction of useful information[1]. Meanwhile, the introduction of machine learning methods has provided automated solutions for diabetes prediction, through learning and pattern recognition in large-scale medical datasets, thus offering improvements in accuracy and efficiency for diabetes prediction within the medical domain. This synergy between academic research and practical application holds immense potential for the medical field, opening new opportunities for the improvement of diagnosis and treatment for diabetes patients. This study proposes a method for diabetes prediction using a Support Vector Machine (SVM) model based on hospital health check-up data. SVM, a type of supervised learning method, is employed for detection, classification, or regression tasks[2]. In comparison to traditional prediction methods, SVM can overcome their limitations and is particularly adept at capturing complex nonlinear relationships. As a potent machine learning algorithm, SVM is capable of handling high-dimensional data and excels in nonlinear data modeling. By leveraging the advantages of SVM, we can predict the likelihood of patients developing diabetes based on various diagnostic indicators, thereby reducing the severity of the disease and the risk of complications. This research provides valuable insights for the management and treatment of diabetes.

2. Literature review

2.1. SVM Model

With the deepening application of machine learning in the medical sector, Support Vector Machines (SVM) have garnered widespread attention in diabetes prediction research due to their superior generalization capabilities and classification accuracy. SVM is a supervised learning method well-suited for binary classification problems, which can be extended to multi-class issues. It operates by identifying the optimal hyperplane in feature space to maximize the margin between disparate categories, thus achieving its classification objective.

The stability and accuracy of SVM are often highlighted in the literature. For instance, studies by Guo Jindan et al.[3] have demonstrated that SVM outperforms other algorithms such as decision trees and random forests in terms of stability and accuracy when applied to diabetes datasets. A key advantage of SVM is its performance with small sample datasets, which is particularly crucial in the medical field where large samples for certain diseases are often unattainable.

The selection of the kernel function is one of the critical components for the successful application of SVM. Through appropriate kernel functions, such as the Radial Basis Function (RBF) or polynomial kernels, SVM can address nonlinear problems, which is vital for disease prediction. However, the choice of kernel functions and parameters often requires expertise and may need different configurations across various datasets[4].

Despite SVM's impressive performance in numerous studies, it has some drawbacks. The tuning process can be complex and computationally intensive, especially with large datasets. Additionally, SVM may not be as effective as some emerging deep learning models in extracting complex features and handling a substantial number of features.

When contrasted with other models such as deep neural networks and ensemble learning algorithms, SVM may not always be the best choice in certain studies. For example, research by Zhou Leming et al.[5] suggests that while XGBoost has advantages in data preprocessing and missing value handling, SVM remains a powerful contender when dealing with datasets with clear boundaries. Hence, future research might explore how to further enhance SVM's performance in disease prediction through algorithm fusion[6] and feature engineering.

2.2. Disease Prediction

Disease prediction is one of the key tasks in the healthcare sector, aiming to prognosticate potential diseases in patients by analyzing clinical data. With the advancement of machine learning technology, disease prediction models are playing an increasingly significant role in improving diagnostic accuracy and patient management. Recent studies have seen scholars employing modern artificial intelligence techniques, particularly machine learning algorithms, to enhance the prediction accuracy of diabetes and its complications.

Advanced algorithms like XGBoost and LightGBM, used by Zhou Leming et al., have shown efficient predictive capabilities with high accuracy when processing large clinical datasets. This indicates that ensemble learning methods hold considerable potential in medical prediction models, though the study was limited to specific case combinations and may not apply to a broader range of diseases. The research by Guo Jindan et al. focused on the performance comparison of common algorithms in predicting type 2 diabetes risk, where the Logistic Regression (LR) method demonstrated the best performance in terms of accuracy and stability, highlighting the effectiveness of traditional statistical methods in certain scenarios.

The work of Xu He et al.[7] emphasizes enhancing model interpretability, which is essential for fostering trust and understanding among medical professionals. By integrating knowledge representation vectors into deep learning models, they offered new perspectives for interpreting the predictions of complex models. While the model improved interpretability, it may also increase complexity due to the introduction of additional knowledge representations. Thus, each model

possesses its unique features and limitations, and the selection of an appropriate model should be based on the specific application context and data characteristics.

These studies suggest that future disease prediction models should focus on improving model accuracy and generalization, enhancing interpretability, and optimizing algorithms to accommodate datasets of varying scales. Furthermore, interdisciplinary collaboration, such as the close cooperation between medical experts and data scientists, will play a pivotal role in constructing more precise and pragmatic disease prediction models.

3. Preprocessing of data

3.1. Data source

The data employed in this study originates from processed medical examination records provided by a certain hospital. The dataset encompasses various parameters including gender, age, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), as well as the presence or absence of diabetes mellitus among other fields.

3.2. Data preprocessing

In consideration of the fact that certain fields within the health examination dataset do not correlate with the etiology of diabetes mellitus, irrelevant data such as minor item names were excluded from the analysis. Consequently, fifteen attributes including age, total cholesterol, and very low-density lipoprotein cholesterol (VLDL-C) were retained as the initial attributes of the health examination dataset. Regarding missing values within the dataset, this study implemented mean imputation for continuous variables and mode imputation for binary categorical attributes. Furthermore, relevant variables were mapped to facilitate the transformation of categorical data into numerical format, with the mapping outcomes detailed in Table 1.

Table 1. Variable mapping tables

Physical Examination Variable Name	Assignment
Gender	0=Male, 1=Female
History of Hypertension	0=None, 1=Present
BMI	0=≤23.9, 1=24.0-27.9, 2=≥28.0
Fasting Blood Glucose	0=≥7.0mmol/L, 1=<7.0mmol/L

Upon visual analysis of the physical examination results, several disparities were discerned between diabetic individuals and their non-diabetic counterparts. The visualization of the examination outcomes is depicted as shown below. Figure 1 indicates a positive correlation between Body Mass Index (BMI) and the prevalence of diabetes, while Figure 2 reflects a marginally higher proportion of healthy females compared to males. According to Figures 3 and 4 , there is a distinct age distribution disparity between diabetic and non-diabetic individuals, with a lower incidence of diabetes in the younger population. Concurrently, the peak distribution of High-Density Lipoprotein Cholesterol (HDL) is slightly reduced in diabetics, whereas the peak for Low-Density Lipoprotein Cholesterol (LDL) is somewhat elevated, which may suggest the involvement of HDL and LDL levels in the pathogenesis of diabetes.

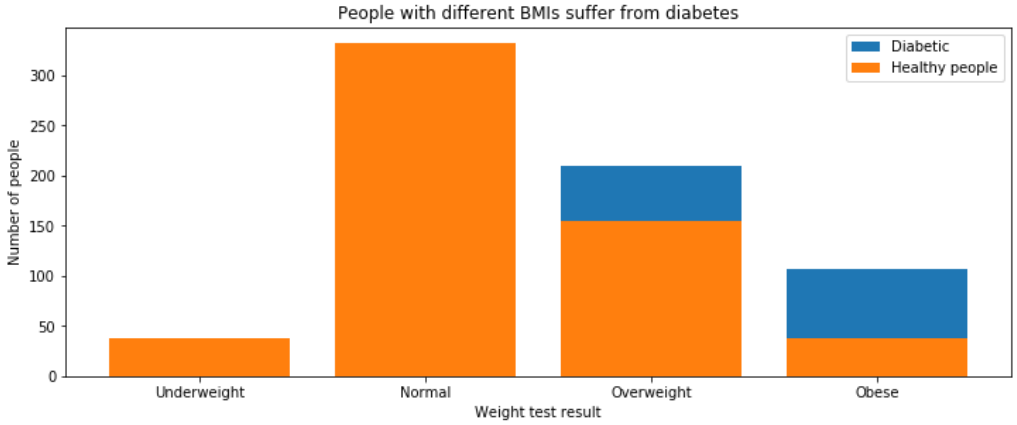

Figure 1. People with different BMIs suffer from diabetes

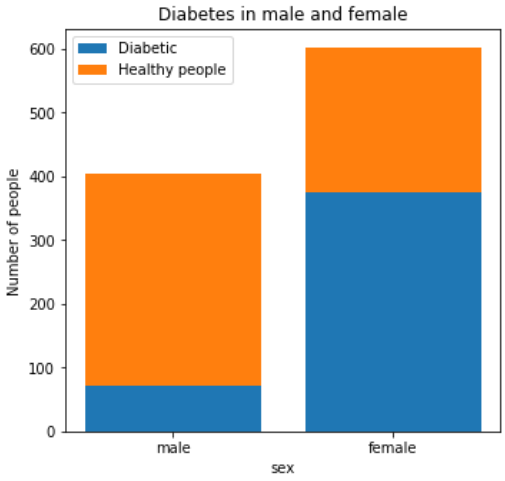

Figure 2. Diabetes in male and female

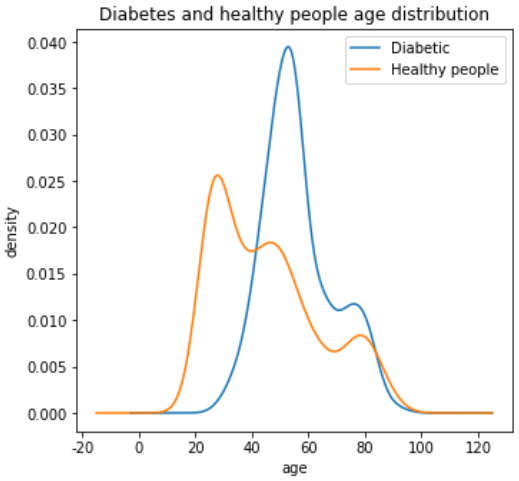

Figure 3. age distribution

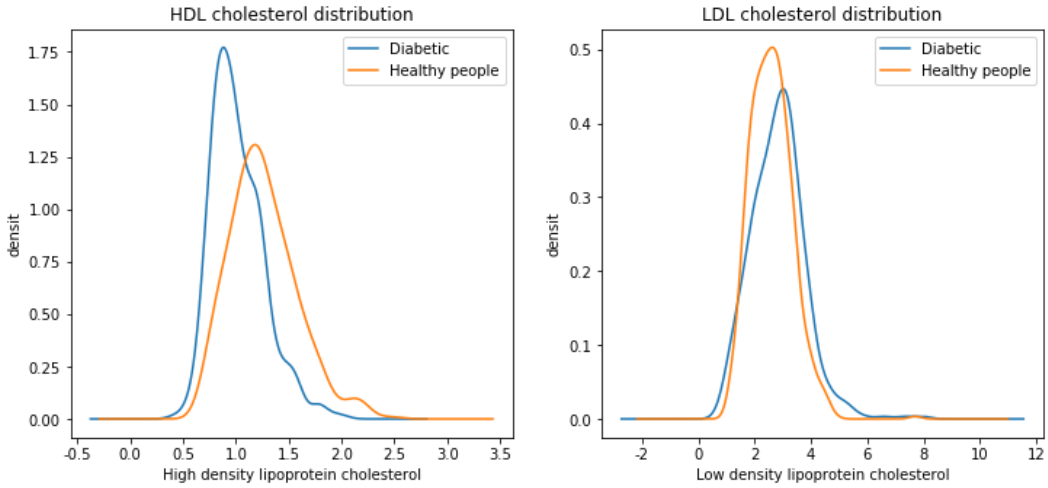

Figure 4. HDL and LDL cholesterol distribution

3.3. Feature Selection for Data

3.3.1 Data Correlation and Multicollinearity Analysis

Irrespective of whether regression algorithms (utilized for numerical predictions) or classification algorithms (employed for category predictions) are applied, it is imperative that features exhibit relevance to the target variable. Should a feature fail to demonstrate a correlation with the target variable, it emerges as a prime candidate for exclusion. Hence, the initial step involves the computation of the correlation amongst the various data attributes. The attribute heatmap is illustrated in Figure 5

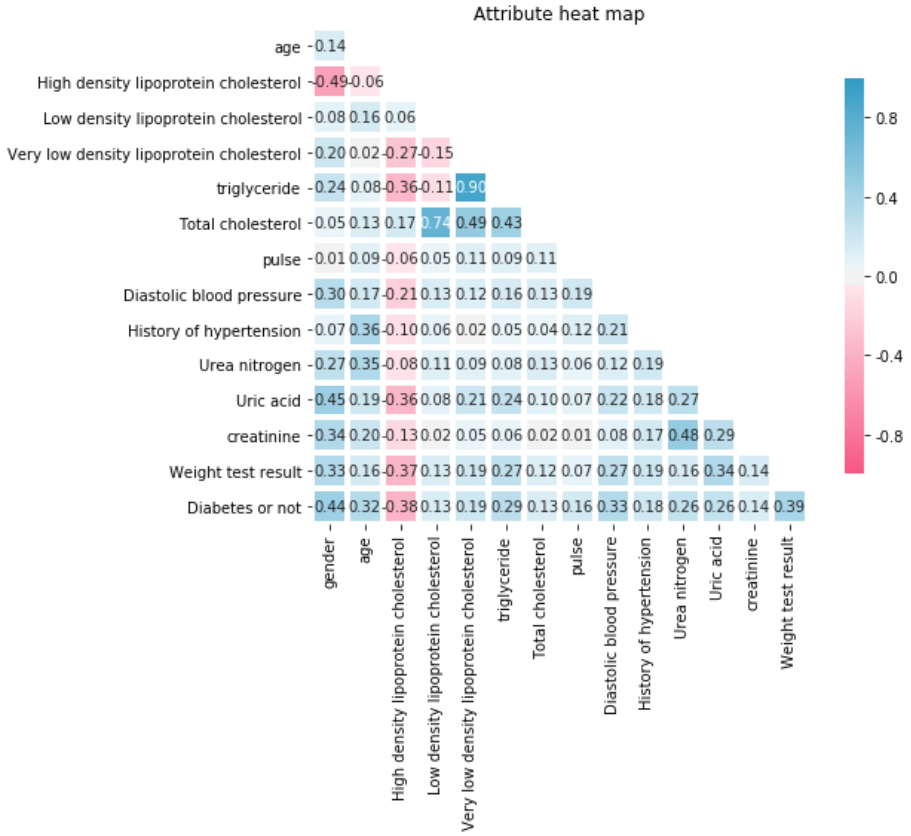

Figure 5. Attribute Heatmap

Multicollinearity arises when there is a correlation between any two features. In machine learning, it is anticipated that each feature should be independent of others, meaning they should not exhibit collinearity. In this study, the Variance Inflation Factor (VIF) is employed to measure multicollinearity among variables, with the computational formula as follows:

$$VIF = \frac{1}{1-R_i^2} \tag{1}$$

Herein, R_i denotes the negative correlation coefficient obtained from the regression analysis of the independent variable against the remaining independent variables.

The results obtained are presented in Table 2. A high VIF for a feature indicates that it is correlated with one or more other features.

Table 2. VIF Values of Attributes

Attributes	VIF
Total Cholesterol	5303.8416
Low-Density Lipoprotein Cholesterol	1751.0207
Very Low-Density Lipoprotein Cholesterol	331.8979
High-Density Lipoprotein Cholesterol	321.8091
Uric Acid	20.3528
Blood Urea Nitrogen	17.9157
Pulse	10.9725
Triglycerides	9.8264
Creatinine	8.2770
Body Weight Examination Result	6.9539
Diastolic Blood Pressure	6.8978
Age	5.1422
Gender	4.2683
History of Hypertension	1.4382

3.3.2 Recursive Feature Elimination and Regularization

In the domain of feature selection, a prevalent method employed is Recursive Feature Elimination (RFE). RFE conducts feature selection by iteratively removing attributes and constructing a model on the remaining features. The fundamental concept underpinning RFE is to leverage the accuracy of the model to evaluate the contribution of each feature (or combination of features) to the predictive outcome.

The implementation steps of the RFE algorithm are as follows:

1. Input the original feature set $Q = (Q_1, Q_2, \dots, Q_m)$

2. Initialize the target feature set $Q^* = q$

3. Train a Support Vector Machine (SVM) based on Q and obtain the weights of all features, square the weights, and then rank the features in descending order.

4. Iteratively execute the steps, removing one attribute at a time, until only one attribute remains, updating Q^* accordingly.

5. If the model fit of the resulting target feature set Q^* does not show any further improvement, and the Mean Squared Error (MSE) does not continue to decrease, the algorithm terminates. The n features thus selected are the n chosen features. If not, return to step 3.

During the feature selection process of Recursive Feature Elimination (RFE), regularization can be utilized to mitigate overfitting. When there is an abundance of features, regularization techniques can shrink the coefficients of features (L2 regularization) or set some feature coefficients to zero (L1 regularization), thereby exerting control over the features. Upon this foundation, RFE refines the selection by iteratively removing features and constructing models, further filtering out features that make significant contributions to the prediction results.

Ultimately, through the above steps, the attributes of age, high-density lipoprotein cholesterol, diastolic blood pressure, pulse, blood urea nitrogen, body weight examination results, gender, and triglycerides are selected as the input for the subsequent SVM model.

4. Model Construction

4.1. Introduction to the Fundamental Concepts of the SVM

The Support Vector Machine (SVM) model is capable of mapping data into a high-dimensional feature space and identifying the optimal separating hyperplane within this feature space to address classification problems. SVMs are adept at overcoming the common issues of underfitting and overfitting that plague most models, requiring fewer samples, offering rapid training, and possessing strong generalization capabilities, exhibiting commendable performance across various scenarios[8]. Given that the sample size in this study is limited to 1006 instances, the SVM model is particularly well-suited for addressing small-sample problems. The eight indicators selected above serve as independent variables, with the presence of diabetes being the dependent variable Y, input into the SVM model.

4.2. Establishment of the Support Vector Machine Model

4.2.1 Model Construction

(1) Normalization of Data

The a priori information contained within the sample data, which includes both training and testing datasets, directly influences the performance of the optimized classifier and the experimental outcomes with the test data. Consequently, it is essential to preprocess the sample data. The objective of preprocessing is to enhance the separability of the data to a reasonable level, thereby ensuring comparability among data of different scales and magnitudes. Furthermore, a normalized data matrix can accelerate the computational speed of the model and effectively circumvent numerical difficulties during calculation. In this study, linear range transformation is employed for data preprocessing. The

preprocessing formula is shown as equation (2), where x_i represents the original sample data, and x_i' denotes the new data obtained through linear range transformation.

$$x_i' = \frac{x_i - min_i}{max_i - min_i} \tag{2}$$

(2) Selection of Kernel Function

The kernel function is pivotal in constructing the optimal classification hyperplane, serving the purpose of establishing a non-linear mapping relationship between the input space, constituted by the high-dimensional training data matrix, and the Hilbert feature space, thereby facilitating the resolution of convex optimization problems within the Hilbert space. Kernel functions are generally categorized into linear and non-linear types, with the most commonly utilized non-linear kernels including polynomial, Gaussian radial basis function (RBF), and multilayer perceptron kernels[9]. When SVMs are applied to classification tasks, the selection of the kernel function and the determination of its parameters are critical. The primary types of kernel functions include linear (LINEAR), polynomial (POLY), Gaussian radial basis function (RBF), and sigmoidal (SIGMOID)—each associated with neuronal non-linear effects[10]. This study employs the Gaussian radial basis function kernel, as denoted by equation (3). This kernel function is capable of mapping non-linear problems to a higher-dimensional feature space, thereby simplifying the data separation in the new feature space. It performs effectively with both large and small sample sizes and exhibits robust resistance to noise present in the data.

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \sigma > 0 \tag{3}$$

(3) Parameter Optimization

The optimization of parameters for the SVM model is commonly carried out via cross-validation and grid search techniques. In this study, grid search was employed for the SVM parameter optimization process. Grid search operates by defining a set of candidate parameter values, followed by training and evaluating the model for each parameter combination to identify the optimal set. Under each parameter set, the model is assessed using cross-validatio, with the evaluation metric being the area under the ROC curve (roc_auc), as designated by the scoring parameter. The specific steps are as follows:

i. Define the parameter space: Initially, a set of candidate parameter values must be defined for exploration, which includes C and $gamma$.

ii. Create the model: Instantiate a basic SVM model object (model_svm).

iii. Grid Search: Employ the GridSearchCV class to create a grid search object, incorporating the model object, parameter space, number of folds for cross-validation, and the scoring metric.

iv. Execute the search: Invoke the `fit` function of the grid search object, inputting the training data set (X_train_minmax and train_y) to commence the parameter search process.

v. Evaluate results: Upon completion of the search process, the best parameter combination can be retrieved via the best_params_attribute, and the corresponding best score through the best_score_attribute.

5. Solution of the Model

5.1. Evaluation Metrics

This study employs Precision and Recall as the evaluation metrics for the SVM model, as depicted in equations (4) and (5):

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Precision and recall, in certain instances, particularly within certain medical contexts, are not sufficient to provide a comprehensive assessment of a machine learning model's predictive capability and efficiency. In some cases, they may lead to serious biases in model evaluation, failing to accurately measure the effectiveness of a classifier and potentially misleading those who use and evaluate the model. To address this issue, a common and relatively straightforward method is the utilization of the F-Measure, which involves evaluating a model's predictive performance by calculating the F1-Score.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

The ROC curve facilitates the easy identification of a classifier's ability to recognize samples at a given threshold. ROC analysis is conducted by plotting the curve on a two-dimensional plane equipped with two coordinate axes. The horizontal axis (x-axis) represents the False Positive Rate (FPR), which indicates the proportion of actual negative samples among those predicted as negative by the trained classifier relative to all actual negative samples; the vertical axis (y-axis) reflects the True Positive Rate (TPR), also known as the Recall, which denotes the proportion of actual positive samples among those predicted as positive by the trained classifier relative to all actual positive samples.

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

5.2. Analysis of Model Resolution

This paper utilizes Python as the execution environment for the project, employing the sklearn library within Python to establish the SVM predictive model. The model inputs are the aforementioned screened attributes: age, high-density lipoprotein cholesterol, diastolic blood pressure, pulse, blood urea nitrogen, body weight examination results, gender, and triglycerides, encompassing a total of eight characteristics. The output of the model's effectiveness is presented as shown in Table 3.

Table 3. SVM model results

F1-Score	Precision	Recall
0.7968	0.7786	0.816

Utilizing an SVM model for diabetes prediction yields an F1 score of approximately 0.797, which is relatively high, indicating that the SVM model achieves commendable comprehensive performance in diabetes predictions. It demonstrates a good balance between prediction accuracy and recall. Consequently, this model can offer higher accuracy and a lower error rate in distinguishing between diabetic and non-diabetic patients.

The accuracy and recall rates are approximately 0.779 and 0.816, respectively, signifying the model's capability to accurately identify patients and its high sensitivity. It effectively captures actual diabetic patients, classifying them correctly, while ensuring that non-diabetic individuals receive an accurate negative diagnosis. This is of significant importance for researchers and physicians in the early diagnosis and intervention of diabetes, aiding in the provision of timely treatment and management.

Qualitative analysis from Figures 6 and 7 suggests that the ROC curve is well above the diagonal line, achieving a high TPR at a low FPR, indicating favorable model performance. Quantitative analysis reveals that the Area Under the Curve (AUC) value is 0.857, which is greater than 0.8, reflecting the overall good performance of the model.

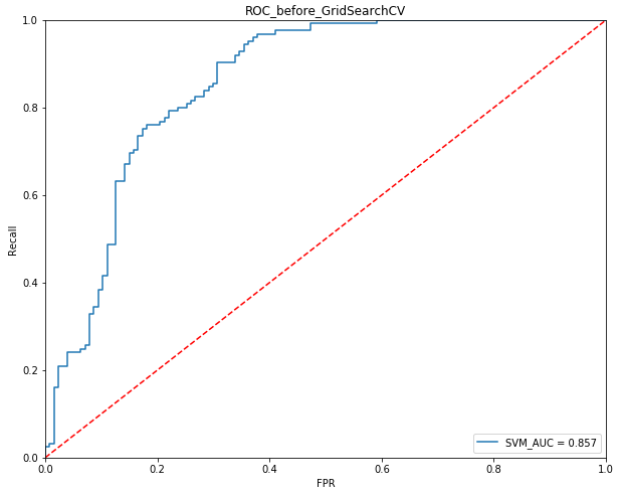

Figure 6. Pre-tuning ROC Curve

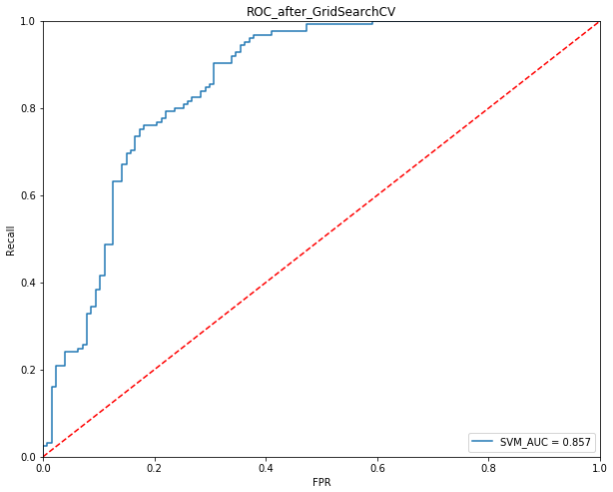

Figure 7. Post-tuning ROC Curve

Although the SVM model exhibits high overall performance on this dataset, optimization of the model's classification accuracy, generalization ability, and robustness is necessary to achieve improved predictive performance. A grid search was employed to find the optimal C and gamma values for the SVM model, resulting in a C value of 100 and a gamma of 0.3. The outcomes post-parameter tuning, as applied to the original SVM model, are displayed in Table 4.

Table 4. SVM model results

F1-Score	Precision	Recall
0.8106	0.7697	0.856

The comparison suggests that the model's performance post-tuning is generally slightly superior to that before tuning.

5.3. Interpretation of SVM Model Application Significance

Diabetes is a common and severe chronic disease where early diagnosis and prevention are crucial for the patient's health. Utilizing an SVM model for diabetes prediction can assist doctors and researchers in identifying individuals at risk of diabetes at an early stage. By taking preemptive measures such as lifestyle changes or medication, the risk of diabetes can be mitigated, thereby enhancing the quality of life for patients.

6. Conclusion

This paper has explored a diabetes prediction method based on hospital physical examination data through the use of a Support Vector Machine (SVM) model. By learning from and recognizing patterns in large-scale medical data, it is possible to precisely predict the likelihood of a patient developing diabetes, which is of significant importance for early risk prediction and intervention. Furthermore, this paper leverages the advantages of SVM, which can handle high-dimensional data and capture complex nonlinear relationships. Compared to traditional prediction methods, SVM overcomes their limitations and enhances prediction accuracy and reliability, successfully forecasting the risk of diabetes development.

However, the limited volume of physical examination data, with only 1006 analyzable records, is insufficient for robust diabetes prediction, and the representativeness of the analyzed results is not strong. The lack of time-series data makes the prediction less accurate since such data better reflects an individual's health status. Therefore, future research could incorporate time-series data, further refine the SVM model to improve predictive performance, and combine other machine learning algorithms and techniques to further perfect the prediction and treatment methods for diabetes.

References

[1] G. Swapna, R. Vinayakumar, K.P. Soman, Soman KP diabetes detection using deep learning algorithms, ICT Express 4 (4) (2018): 243–246.

[2] S. Chidambaranathan, A. Radhika, V.V. Priya, S.K. Mohan, M.G. Gireeshan,Optimal SVM based brain tumor MRI image classication in cloud internet of medical things, in:Cognitive Internet of Medical Things for Smart Healthcare,Springer, Cham, 2021, pp. 87–103.

[3] Guo Jindan, Gao Yanyan, Gao Huailin, et al. Comparative Study on the Performance of Type 2 Diabetes Risk Prediction Models. China Biotechnology, 2023, 43(11): 35-42.

[4] Zhen Tong, Fan Yanfeng. Research on Corporate Credit Risk Assessment Based on Support Vector Machine. Microelectronics & Computer, 2003, 23(5): 136-139.

[5] Zhou Leming, Shang Mingsheng, Wang Yonghong, et al. Study on Diabetes Prediction Based on Artificial Intelligence. Journal of Chongqing Medical University [J/OL]: 1-4 [2024-01-03].

[6] Ou Quanhong, Shi Lifei, Cheng Feiyan, et al. Study on the Diagnosis of Lung and Liver Cancer Based on Serum Infrared Spectroscopy Combined with SVM Algorithm. Spectroscopy and Spectral Analysis, 2023, 43(S1): 57-58.

[7] Xu He, Zheng Qunli, Xie Zuoling, et al. Interpretable Deep Learning Models Based on Knowledge Representation Vectors and Their Applications in Disease Prediction. Data Acquisition and Processing, 2023, 38(04): 777-791.

[8] Yang Meitao, Wang Yanding, Li Zhiqiang, et al. Application of ARIMA-SVM Combined Model in Predicting the Incidence Trend of Pulmonary Tuberculosis. Modern Preventive Medicine, 2023, 50(11): 1921-1926.

[9] Hu Haiqing, Zhang Lang, Zhang Daohong. Research on SME Credit Risk Assessment from the Perspective of Supply Chain Finance—A Comparative Study Based on SVM and BP Neural Network. Management Review, 2012, 24(11): 70-80.

[10] Chen Yuanfeng, Ma Xiyuan, Cheng Kai, et al. New Energy Ultra-Short-Term Power Forecasting Based on Meteorological Feature Selection and SVM Model Parameter Optimization. Acta Energiae Solaris Sinica, 2023, 44(12): 568-576.