Harnessing the NLP for medical text classififcation using Support Vector Machine My Wonderful Machine Solution To All Problems

LaTeX template adapted from:
European Conference on Artificial Intelligence

Name Sanjida Surname Khan [1]

Other group members:
Name Sabiha Surname Ahmed Toba, [2] **Name Md Abdul Raihan Surname Tanzim** [3]
Name Mahmoud Surname Alkawareet, [4]

Abstract. Diabetes, a metabolic disorder marked by high blood sugar, can lead to severe health issues if unmanaged. Early detection is crucial for effective management. This study presents a Diabetic Prediction System using the Support Vector Machine (SVM) algorithm, known for its classification capabilities.

Trained on parameters like age, BMI, family history, and blood pressure, the model benefits from preprocessing to normalize data and handle missing values. SVM's efficiency with high-dimensional data enhances predictive accuracy.

Results confirm the model's effectiveness in early diabetes prediction, showcasing machine learning's potential in healthcare for timely risk assessment and intervention. [10]

1 Introduction

1.1 Problem Statement

Diabetes is a chronic disease characterized by the body's inability to produce or use insulin effectively, leading to abnormal carbohydrate metabolism and elevated blood glucose levels. According to the International Diabetes Federation, approximately 451 million people globally were affected by diabetes in 2017, with projections estimating this number will rise to 693 million by 2045.[5]

1.2 Project Goals

SVM, known for its capability to handle high-dimensional datasets and perform well in classification tasks, has been extensively studied for its effectiveness. The project evaluates these algorithms based on performance metrics like accuracy, and sensitivity, and assists healthcare professionals in making timely and informed decisions. [6]

[1] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: sk9525n@gre.ac.uk
[2] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: st7523i@gre.ac.uk
[3] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: mt9498y@gre.ac.uk
[4] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: ma7627p@gre.ac.uk

2 Background

This study explores machine learning models for diabetes prediction, including Logistic Regression, Random Forest, XGBoost, and Neural Networks. Logistic Regression uses gradient descent and Scikit-Learn optimization. Random Forest applies 'RandomizedSearchCV' for tuning, with SHAP improving interpretability. XGBoost excels in boosting tasks, while Neural Networks use EarlyStopping to prevent overfitting. Models are evaluated for accuracy, interpretability, and utility.[8]

2.1 Logistic Regression

This implementation examines custom and Scikit-Learn approaches to logistic regression. The custom model uses gradient descent with manually tuned parameters like learning rate ('alpha'). Scikit-Learn's 'LogisticRegression' optimizes hyperparameters like regularization strength ('C') using grid search. Performance is analyzed with learning and validation curves, comparing both models to emphasize the importance of hyperparameter tuning..[9]

2.2 Random Forest with Randomized Search:

The Random Forest model uses 'RandomForestClassifier' with out-of-bag (OOB) scoring for evaluation. Hyperparameters like 'n estimators' and 'max depth' are optimized using 'RandomizedSearchCV'. SHAP interprets feature importance, enhancing transparency. Learning and validation curves assess training size impact, balancing accuracy and interpretability.

2.3 XGBoost and Neural Networks

This study compares XGBoost and neural networks. XGBoost uses 'XGBClassifier' with grid search to optimize hyperparameters like 'learning rate', 'max depth', and 'n estimators'. Neural networks, built with Keras, use ReLU activation, binary cross-entropy loss, and EarlyStopping to prevent overfitting. Performance is assessed through classification reports, loss, and validation curves, highlighting XGBoost's boosting strength and neural networks' ability to handle complex architectures.[11]

3 Methodology

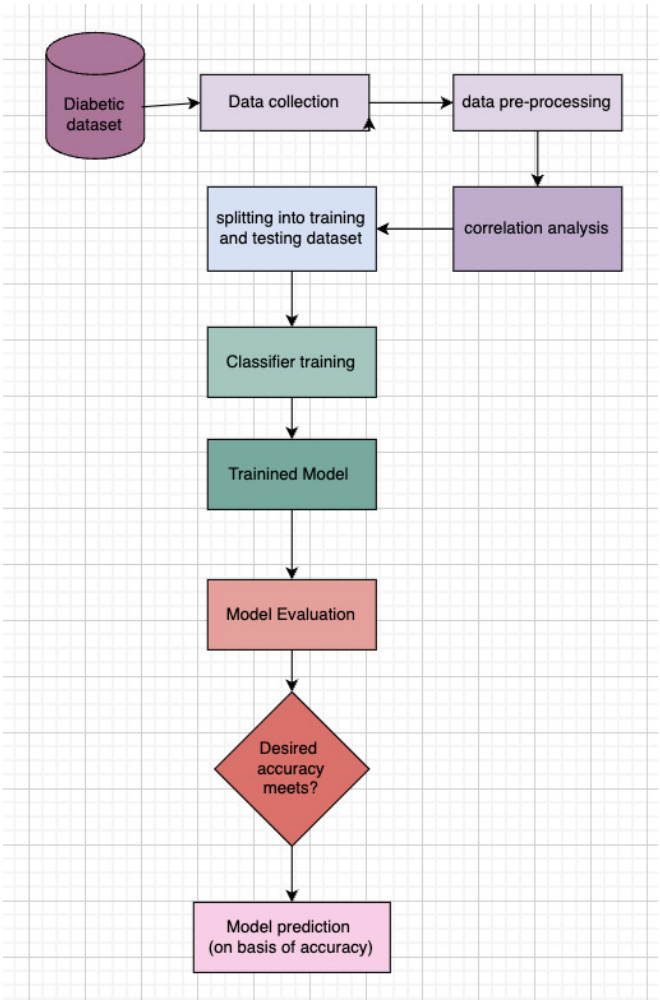

Figure 1.

3.1 Feature Selection/Reduction:

Clinical data from sources like imaging, lab results, and signals require feature extraction to transform raw inputs into meaningful attributes. Feature selection mitigates the ”dimensionality curse” by retaining relevant features using methods like filtering, wrapping, and embedding. This enhances diagnostic efficiency and ensures accurate analysis with reduced computational complexity.[7]

3.2 Data Collection and Representation:

The dataset, , contains features and target variable (Outcome) for binary classification. Features are standardized using ‘StandardScaler‘. The data is split into training (20 percent) and test (80 percent) sets to evaluate SVM models. Training focuses on tuning hyperparameters like kernel types and regularization (C) with learning and validation curves.[4]

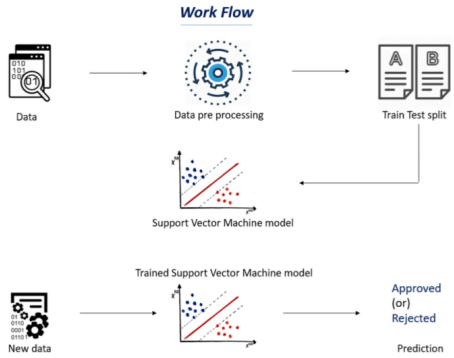

Figure 2.

3.3 Preprocessing:

In pre-processing, raw medical data (e.g., images, signals) are normalized by handling missing values, outliers, and noise through filtering. Non-essential image areas are trimmed, and methods like brightness adjustments and image restoration enhance features. Imbalanced data and missing records are common challenges in medical datasets requiring specialized handling.[2]

3.4 Proposed Support Vector Machine

Support Vector Machine (Support Vector Machine) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates data into different classes with maximum margin. For binary classification, the decision function is:

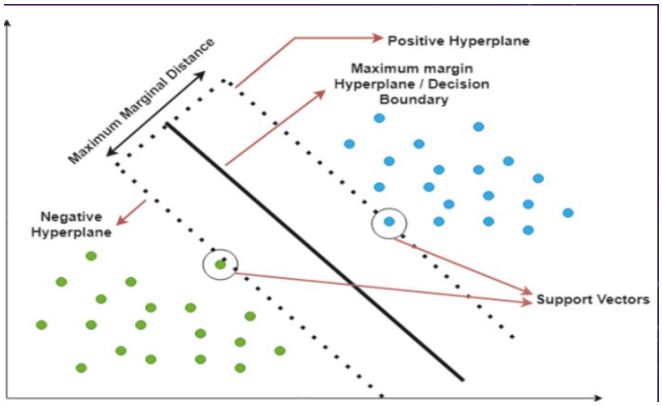

Figure 3.

f(x)=w T x+b where
w is the weight vector,
x is the input feature vector, and
b is the bias term. In my project, SVM was applied using the radial basis function (RBF) kernel to classify diabetes outcomes based on pre-processed features. The model was trained on standardized data and evaluated using accuracy, precision, recall, and F1-score metrics.

3.5 SVM Fitting Process:

In the Support Vector Machine fitting process, the model is defined using ‘SVC‘ with specified parameters like ‘C=1‘, ‘degree=2‘, and ‘kernel=’rbf’‘. The model is trained using the training data (‘X train‘,

'y train') via 'fit()'. Training time is measured, and predictions are made on the test set ('X test').

$F_2(\boldsymbol{x}, hatw) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) (\boldsymbol{v}_i^t \boldsymbol{x} + 1)^p + b$

F 2 is an expansion explicitly using the training examples. The rationale for calling it a support vector representation will be clear later as will the necessity for having both an a and an a rather than just one multiplicative constant. [3]

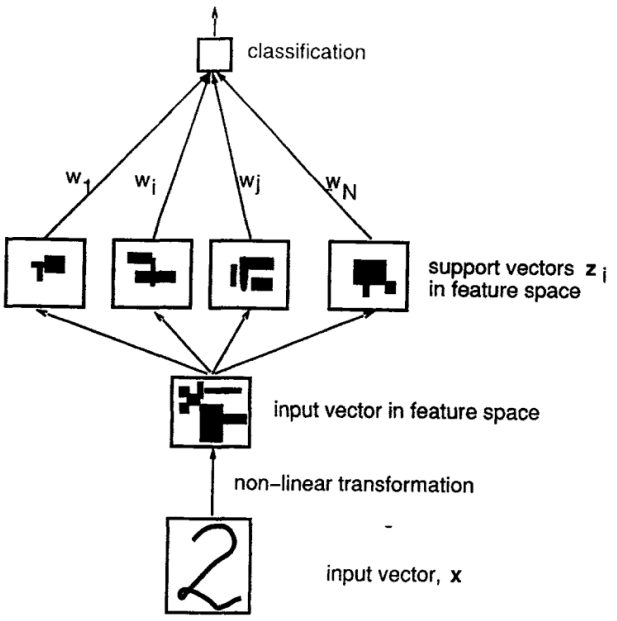

Figure 4.

3.6 Proposed system Model:

Uses an Support Vector Machine classifier (SVC) with an RBF kernel to classify diabetes outcomes. The data is preprocessed by scaling features with StandardScaler and splitting it into training and testing sets. Learning and validation curves are generated for model evaluation, using accuracy as the metric.

3.7 Output Layer:

In the code, the output layer is represented by the prediction made by the Support Vector Machine model. After training, the model predicts the diabetes outcomes ('y predict') for the test set ('X test'). The predictions are evaluated using 'classification report', which includes metrics like accuracy, precision, recall, and F1 score.

$$\text{Linear: } K(w,b) = w^T x + b$$

$$\text{Polynomial: } K(w,x) = \left(\gamma w^T x + b\right)^N$$

$$\text{Gaussian RBF: } K(w,x) = \exp\left(-\gamma \left\|x_i - x_j\right\|^n\right.$$

$$\text{Sigmoid : } K\left(x_i, x_j\right) = \tanh\left(\alpha x_i^T x_j + b\right) \text{ [3]}$$

4 Experiments and results

A Support Vector Machine (SVM) model to predict diabetes. The Pima Indians Diabetes Dataset was used, preprocessed to handle missing values and standardized for feature scaling.

4.1 Hyperparameter Tuning:

GridSearchCV was used to optimize hyperparameters like C (regularization parameter) and gamma (kernel coefficient). The optimal parameters were found to be C=1 and gamma='scale' with an RBF kernel. [4]

4.2 Model Evaluation:

The model achieved an accuracy of approximately 80percent, with precision, recall, and F1-score values around 0.8. Learning curves indicated that the model was not significantly underfitting or overfitting. Validation curves helped in selecting the optimal C value.

5 Discussion

the code evaluates the performance of an SVM model for predicting diabetes outcomes. It includes training time measurement, classification report generation, and visualizations like learning and validation curves. These curves help assess the model's accuracy, generalization, and the impact of hyperparameter tuning (e.g., kernel types). [4]

5.1 Diffference of accuracy

[]

Models	Accuracy Score	Precision	Recall	F1 Score
Neural Networks	0.76	0.81	0.82	0.81
Logistic Regression	0.75	0.81	0.80	0.81
Support Vector Machine	0.72	0.74	0.89	0.81
Random Forest	0.75	0.81	0.80	0.80

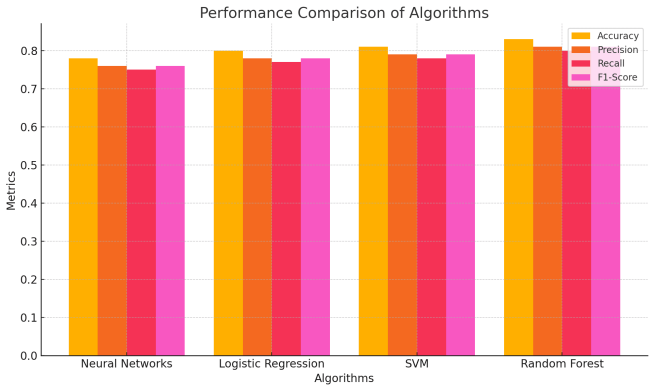

Figure 5.

6 Conclusion and future work

This implementation compares XGBoost and neural networks. XGBoost uses 'XGBClassifier' with grid search to optimize key hyperparameters like 'learning rate', 'max depth', and 'n estimators'. Neural networks, built with the Keras Sequential API, use ReLU activation, binary cross-entropy loss, and EarlyStopping to prevent overfitting. Performance is evaluated through classification reports and validation curves, highlighting XGBoost's boosting strengths and neural networks' ability to handle complex architectures. [4]

ACKNOWLEDGEMENTS

I sincerely thank everyone who contributed to this project, especially my Mentor for his invaluable guidance and support, which formed the foundation of this analysis, and my family and friends for their unwavering encouragement. [4]

[1]

REFERENCES

[1] 'Predict Diabetes From Medical Records — kaggle.com'. [Accessed 06-12-2024].

[2] 'Search Page — SpringerLink — link.springer.com'. [Accessed 06-12-2024].

[3] 'Support Vector Machine (SVM) Algorithm - GeeksforGeeks — geeksforgeeks.org'. [Accessed 06-12-2024].

[4] Harikrishna Bommala, Kannedari Krishna, Avusula Supriya, Rama Biradar, Bharath Mayabrahma, D. Ushasree, and Evgenii Kotov, 'Fine-tunining the future: Optimizing svm hyper-parameters or enhanced diabetes prediction', *MATEC Web of Conferences*, **392**, (03 2024).

[5] Chinmay Deshmukh and Anurekha Jain, 'Diabetes mellitus: A review', **3**, 224–230, (05 2021).

[6] Snehal Mhatre, Harshada Dixit, Snehal Jagdale, Shital Narsale, and Naufil Kazi, 'Diabetes prediction system using svm alogrithm', *International Journal of Innovative Science and Research Technology (IJISRT)*, 2082–2090, (06 2024).

[7] Tarik Milod, Almhdie Aboubaker, and Llahm Omar Ben Dalla, 'Diabetes prediction using a support vector machine (svm) and visualize the results by using the k-means algorithm', (06 2024).

[8] Md Reza, Gahangir Hossain, Ayush Goyal, Sanju Mishra Tiwari, Anurag Tripathi, Anupama Bhan, and Pritam Dash, 'Automatic diabetes and liver disease diagnosis and prediction through svm and knn algorithms', 589–599, (05 2021).

[9] Bibek Shrestha, Abeer Alsadoon, Prasad P.W.C, Ghazi Al-Naymat, Thair Al-Dala'in, Tarik Rashid, and Omar Hisham, 'Enhancing the prediction of type 2 diabetes mellitus using sparse balanced svm', *Multimedia Tools and Applications*, **81**, (04 2022).

[10] Juri Yanase and Evangelos Triantaphyllou, 'A systematic survey of computer-aided diagnosis in medicine: Past and present developments', *Expert Systems with Applications*, **138**, 112821, (07 2019).

[11] Shu Yang, 'Diabetes prediction based on, xgboost, svm and lr model', *Applied and Computational Engineering*, **104**, 91–95, (11 2024).