PERFORMANCE ANALYSIS OF LR AND SVM MODELS FOR DIABETES PREDICTION

PRESENTED BY
VYSHNAVI SANIKOMMU

OUR RESEARCH WORK: HOW HYPERPARAMETER TUNING AND SEARCH STRATEGIES IMPROVE PERFORMANCE OF LR AND SVM MODELS FOR DIABETES PREDICTION

Why this research is important	Diabetes is a lethal disease affecting people worldwide. Early prediction through machine learning models is crucial for preventing diabetes progression by effectively identifying individuals at risk.
What we know and don’t know	Early prediction relies on a model's performance metrics for a specific dataset. However, the influence of hyperparameter tuning and search strategies in enhancing performance and identifying the optimal model remains uncertain.
Our experiment	Implement base models and set up grid search for each model by configuring model-specific hyperparameters and include performance metrics like accuracy, f1-score, confusion matrix, precision and recall.
Our hypothesis	We predict that optimizing hyperparameters through systematic search and tuning will either maintain or enhance performance metrics of each model, leading to the identification of an optimal model for early prediction of diabetes.

DESIGN/METHODS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Study Population	• Diabetic and Pre-diabetic patients • Medical diagnosis centers
Data Collection	Diabetes data collected from Kaggle source • Data consists of 2000 instances and 9 features/attributes • Attributes are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Age, Diabetes Pedigree Function, Outcome
Data Analysis	• Load and visualize the data • Identify dependent and independent features • Identify missing values
Data Preprocessing	• Handling missing values and zero value attributes • Applied simple imputer for zero value attributes with mean
Feature Selection	• Perform feature selection to get quality results • Correlation Matrix

DESIGN/METHODS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Data Normalization	• Process of scaling and transforming numeric features to a standard scale or distribution • Ensure all features contribute equally to model
Data Splitting	• Split data into training and testing sets with 80:20 ratio • Split using train_test_split() method • 1600 instances in training and 400 instances in testing
Models	Chosen two models for diabetes prediction • Logistic Regression • Support Vector Machine
Hyperparameter Tuning	Chosen GridSearchCV search strategy for tunning two models • Create param_grid specific to each model • Specify cross-validation with N folds • Base models considered as estimators
Performance Evaluation	Performance comparison • Accuracy, Precision, Recall, F1-score, Confusion matrix • Classification report

DATA/RESULTS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Data Analysis	• All features are numeric and no null values • Two unique values for target variable – Outcome, 0 (non-diabetic) and 1 (diabetic) • Describe the min, max, variance and count for each feature • There are 1316 non-diabetic and 684 diabetic instances
Data Preprocessing	• Handled features with values as 0 using simple imputer • Imputer replace zero values with mean corresponding to feature
Feature Selection	• Identify the correlation between input and output features • Calculate correlation matrix • Dropped 2 features Blood Pressure and Diabetes Pedigree Function based on cut-off 0.2

DATA/RESULTS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Data Normalization	• Normalize the input data using Standard Scalar • Ensure features contribute equally to the learning process, preventing certain features from dominating others
Logistic Regression	• Implement base model with default params with C = 1.0 • LR achieve 77% and 79% accuracy on train and test sets • LR achieves 0.75 precision, 0.58 recall and 0.65 F1-score on test set
Support Vector Machine	• Implement base model with default params with C = 1.0, Kernel = rbf • SVM achieve 83% and 82% accuracy on train and test sets • SVM achieves 0.77 precision, 0.66 recall and 0.71 F1-score on test set

DATA/RESULTS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Tuning LR model with GridSearchCV	• Implement model with param grid, estimator and cross-validation • Meta parameters – regularization strength (C), solver, penalty (l1, l2, elastic net) • LR achieve 77% and 81% accuracy on train and test sets • LR achieves 0.80 precision, 0.57 recall and 0.67 F1-score on test set
Tuning SVM model with GridSearchCV	• Implement model with param grid, estimator and cross-validation • Meta parameters – regularization parameter (C), kernel, gamma, degree • SVM achieve 85% and 82% accuracy on train and test sets • SVM achieves 0.78 precision, 0.66 recall and 0.72 F1-score on test set
Performance Evaluation	• Accuracy • Precision, Recall and F1-score • Confusion Matrix • Classification Report

CONCLUSION IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

LR and SVM demonstrated promising results for diabetes prediction

Hyperparameter tuning with GridSearchCV further improved the model’s performance

SVM outperforms LR by achieving highest accuracy of 82% on training and testing sets

SVM achieves good precision, recall and F1-score reflects fair balance between precision and recall

LR achieves good precision, lower recall and F1-score suggests the model is moderately balanced

The optimized models achieved higher accuracy scores highlighting the effectiveness of hyperparameter tuning in enhancing predictive capabilities.

NEXT STEPS IN STUDY OF HYPERPARAMETER TUNING WITH SEARCH STRATEGIES TO ENHANCE MODEL PERFORMANCE FOR DIABETES PREDICTION

Data Collection	• Expansion of dataset • Improves models generalizability and robustness • Reduce bias between classes
Class Imbalance	• Model bias towards majority class • Techniques like undersampling, oversampling, SMOTE • Improves recall by increasing false negatives
Diverse ML models	• Use advanced ML and deep learning models • Capture patterns in the dataset which traditional ML would miss • Improves model performance
Search strategies	• Use efficient tuning processes • RandomizedSearchCV and Bayesian optimization strategies • Achieve higher accuracy and reliability

QUESTIONS

THANK YOU