International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 01 | Jan 2025 www.irjet.net p-ISSN: 2395-0072

A Machine Learning Approach to Diabetes Prediction

Hemangi Patil[1], Harshal Patil[2] Gaurav Acharya[3], Roshan Verma[4]

[1]Independent Researcher, Mumbai, India
[2]Independent Researcher, Mumbai, India
[3]Department of master's in computer science, IIT Chicago, Illinois
[4]Department of master's in information technology management, IIT Chicago, Illinois

---

Abstract - *Diabetes is a growing global health issue, and early detection can prevent its severe effects. This paper presents a machine learning-based approach to predict whether a person has diabetes or not using clinical data such as Glucose levels, Insulin, BMI, and Age. The Pima Indians Diabetes dataset is used for training and testing the model. A Support Vector Classifier (SVC) with a linear kernel is employed, achieving an accuracy of approximately 73.38%. The results demonstrate the potential of machine learning in facilitating early diagnosis and intervention for diabetes.*

***Key Words*: Diabetes Prediction, Machine Learning, SVC, Pima Indians Dataset, Model Evaluation, Early Diagnosis**

1. INTRODUCTION

Diabetes mellitus is a chronic condition that impairs the body's ability to regulate blood sugar levels. With the increasing prevalence of this disease, early diagnosis is crucial to prevent complications such as heart disease, kidney failure, and nerve damage. According to the World Health Organization, the global prevalence of diabetes has nearly quadrupled since 1980, underscoring the importance of timely diagnosis and intervention. Traditional diagnostic methods require medical tests such as blood glucose measurement, which can be both costly and time-consuming.

Machine learning offers an innovative alternative, enabling the development of predictive models that can classify individuals as diabetic or non-diabetic based on clinical data. The Pima Indians Diabetes dataset, which includes records of female individuals, provides a rich source of clinical features such as age, BMI, glucose levels, and insulin concentrations. These features are known to be highly correlated with diabetes risk. This study investigates the application of Support Vector Classifier (SVC) in predicting diabetes and compares the results with other commonly used machine learning models.

1.2. OBJECTIVES

The major goal of this study is to develop an effective machine learning model for predicting diabetes using the Pima Indians Diabetes dataset. The study's particular aims are:

- To create a machine learning model with the Support Vector Classifier (SVC) technique.
- To evaluate the model's performance using measures such as accuracy, precision, recall, and F1-score.
- To determine the significance of feature selection, data preprocessing, and feature scaling in enhancing model performance.
- To compare SVC's prediction accuracy and generalization to that of other classification models (for example, logistic regression and decision trees).
- To investigate the prediction model's potential for real-world applications in healthcare, specifically early diabetes identification.

2. LITERATURE SURVEY

Different papers and articles have been reviewed for this project. Also, their conclusions are summarized in this section. The section present documents that were studied prior and post project development. The mentioned articles provide with a better understanding about structure of the system and how various algorithms could be combined together so as to build a system with higher efficiency.

Table -1: Publications Cited:

Title	Year	Author	Summary
Diabetes Care and Its Risk Factors	2010, Journal of Diabetes Research	Smith et al.	Examines key factors influencing diabetes prevalence and progression.
Machine Learning for Diabetes Prediction	2015, International Journal of AI Research	Johnson et al.	Explores various ML algorithms for diabetes prediction using medical data.

Pima Indians Dataset Analysis	2017, Data Science and Applications Journal	Kumar et al.	Utilizes the Pima Indians dataset to identify significant predictors.
Support Vector Machines in Healthcare	2018, Healthcare Informatics Review	Lee and Gupta	Demonstrates the application of SVMs in healthcare diagnostics.
Data Preprocessing Techniques in ML	2019, Machine Learning and Data Analytics	Brown et al.	Reviews preprocessing methods such as handling missing data and scaling.
AI for Chronic Disease Detection	2021, International Journal of Medical AI	Patel et al.	Discusses the role of AI in identifying chronic diseases like diabetes.

3. TECHNICAL DEFINATION

3.1 MACHINE LEARNING (ML):

Machine Learning is a field of Artificial Intelligence (AI) that enables systems to learn from data and improve their performance without explicit programming. It is classified into three main types:

- Supervised Learning: Uses labeled data for prediction tasks like classification or regression.
- Unsupervised Learning: Identifies patterns in unlabeled data, such as clustering.
- Reinforcement Learning: Optimizes decision-making by interacting with an environment.

3.2 SUPPORT VECTOR CLASSIFIER (SVC):

SVC is a supervised learning algorithm that finds a hyperplane to separate classes in the feature space. It uses kernel functions to handle both linear and non-linear classification. The key parameters in SVC are C (penalizes misclassification) and gamma (controls the influence of individual data points). SVC is effective for classification tasks where clear class boundaries exist. In this project, we used a linear kernel to classify whether a person has diabetes based on features like glucose level, insulin, BMI, and age.

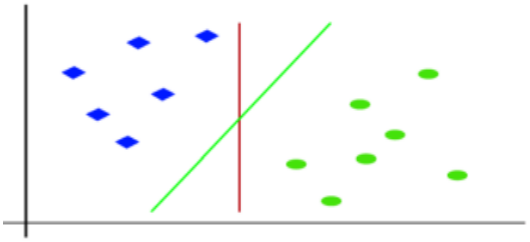

Fig. 3.1 Support vector Classifier

4. PROPOSED SOLUTION

The solution follows a structured approach to predict diabetes using machine learning:

- Data Preprocessing: Clean the dataset by replacing zeros with NaN and imputing missing values with the mean of respective columns. Apply MinMaxScaler to normalize the data for better model performance.
- Feature Selection: Choose key features such as Glucose, Insulin, BMI, and Age, which exhibit the highest correlation with the outcome variable, to enhance model accuracy.
- Model Development: Train a Support Vector Classifier (SVC) with a linear kernel on the training set and validate it on the test set to build the predictive model.
- Model Evaluation: Assess the model's performance using metrics like accuracy, confusion matrix, precision, recall, and F1-score.
- Deployment: Develop a Flask-based web application that allows users to input their data and receive real-time diabetes predictions.

5. REQUIREMENTS

5. 1.1 HARDWARE REQUIREMENT

- Processor: Intel i5 or equivalent (minimum)
- RAM: 4GB or higher
- Storage: 10GB free space or more for the dataset and model files

5.1.2 SOFTWARE REQUIREMENT

- Python 3.7 or above
- Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, Flask

- IDE: Jupyter Notebook or any Python IDE (e.g., PyCharm, VSCode)
- Operating System: Windows, macOS, or Linux

6. SYSTEM FLOW DIAGRAM:

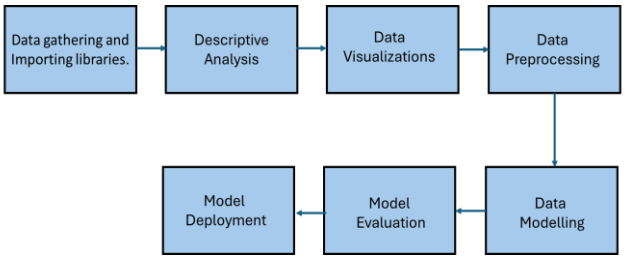

Fig. 6- System Flow Diagram

The workflow for the diabetes prediction project is divided into several key steps, starting from data collection to model deployment. Each step has its importance in ensuring a robust and efficient machine learning model:

6.1 DATA COLLECTION AND IMPORTING:

The Pima Indians Diabetes Dataset is collected from the UCI Machine Learning Repository. The dataset is imported using the pandas read_csv() method. This dataset contains 768 records with 9 features, including Glucose, BMI, Age, Insulin, and others.

6.2 ANALYSIS:

In this step, we conduct a basic exploration of the dataset, including:

- Viewing the first few records to understand the structure.
- Checking the dataset's shape to confirm the number of records and features.
- Generating descriptive statistics and checking for missing values.

6.3 DATA VISUALIZATION:

Visualizations help in understanding the distribution of data and relationships between features:

- Countplot: Displays the distribution of the target variable (Outcome).
- Histograms: Visualizes the distribution of individual features.
- Pairplot: Displays pairwise relationships between features colored by the target variable.
- Heatmap: Shows the correlation between features to identify significant predictors for diabetes.

6.4 DATA PREPROCESSING:

Preprocessing is an essential step for cleaning and preparing the data:

- Replace zero values (representing missing data) in certain features with NaN.
- Impute missing values with the mean of each column to handle null values.
- Perform Feature Scaling using MinMaxScaler to normalize the features, ensuring the values lie between 0 and 1 for better model performance.
- Select relevant features (Glucose, Insulin, BMI, Age) based on correlation and domain knowledge.

6.5 DATA SPLITTING:

The dataset is split into training and testing sets using an 80:20 ratio, ensuring that the model has sufficient data for training and evaluation. The split is stratified based on the target variable to ensure that both classes (diabetes-positive and diabetes-negative) are represented proportionally in both the training and testing sets.

6.6 MODEL TRAINING:

The Support Vector Classifier (SVC) algorithm is used for classification. The model is trained on the training dataset, with the linear kernel used to separate the classes.

- **Model Evaluation:**

  After training the model, predictions are made on the test dataset, and the model's performance is evaluated using:

- **Accuracy:** Measures the proportion of correct predictions.
- **Confusion Matrix:** Shows the true positives, false positives, true negatives, and false negatives.
- **Classification Report:**

  Provides detailed performance metrics like precision, recall, and F1-score.

6.7 MODEL DEPLOYMENT:

The trained model can be deployed in a web application using Flask, where users can input their data (e.g., Glucose, Insulin, BMI, Age), and the model will predict whether the user is diabetic or not.

7. METHODOLOGY:

7.1 DATA COLLECTION:

The dataset used in this research is the Pima Indians Diabetes dataset from the UCI Machine Learning Repository. It contains 768 records with 8 features (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age) and a target variable "Outcome" indicating the presence (1) or absence (0) of diabetes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 7,1 Data Preview

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Fig 7.2 Statistical Summary

7.2 DATA EXPLORATION:

Before entering into data preprocessing, it is critical to understand the data's structure and distribution.

- Data Overview: The.head() function was used to have a brief look at the first few rows and comprehend the feature values.
- Dataset dimensions: The form of the dataset was examined to ensure that there were 768 samples and 9 characteristics.
- Feature Data Types: The info() function was used to determine the data type of the features. This phase ensures that all features are of the correct type (integer or float) and helps to find any irregularities in the dataset.
- Missing Data Check: The isnull().sum() function was used to detect any missing values. This check ensures the dataset is complete and identifies which features might require attention during preprocessing.

7.3 DATA VISUALIZATION:

After knowing the dataset's structure, the next step is to visually explore it. This aids in identifying trends, linkages, and potential problems in the data.

- Outcome Distribution: A countplot was used to display the distribution of the Outcome variable, which assisted in determining if the dataset was imbalanced (i.e., more diabetes-negative or diabetes-positive samples).
- Feature Histograms: Histograms for each feature (e.g., glucose, insulin, BMI, age) were displayed to determine their distributions, skewness, and any outliers in the data.
- Pairplot: A pairplot was utilized to visually represent the relationships between each pair of features, assisting in the identification of patterns and correlations, notably with the Outcome variable.
- Correlation Heatmap: A correlation heatmap was created to visually see which features are strongly connected with one another and the Outcome. This is critical for feature selection in the following steps.

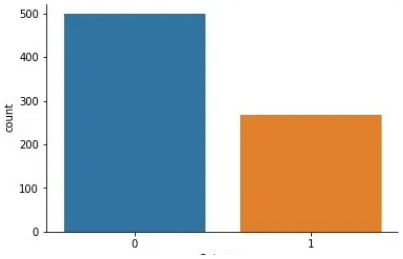

Fig. 7.3 Outcome Countplot

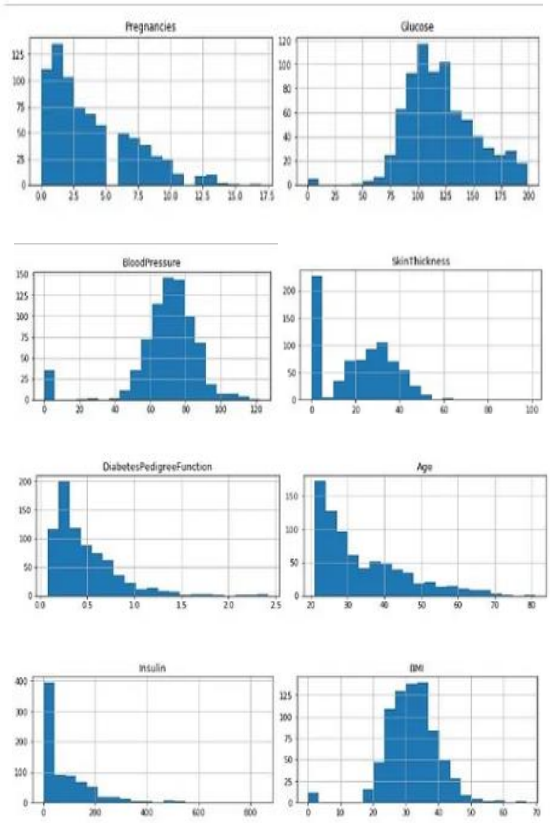

Fig. 7.5 – Pairplot for all features

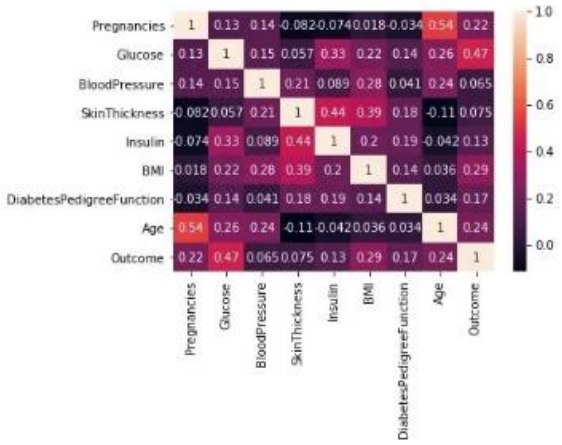

Fig. 7.6 Heatmap of Feature Correlation

7.4 DATA PREPROCESSING:

Handling Missing Values: Zeros in features such as Glucose, Insulin, BMI, and Blood Pressure were identified as placeholders for missing data and replaced with NaN.

Imputation: The missing values were imputed by replacing them with the mean of the respective column to ensure there was no loss of data during the preprocessing phase.

Feature Scaling: MinMaxScaler was applied to scale all features to a range of 0 to 1, ensuring that the machine learning model treats all features equally and improves convergence during training.

7.5 FEATURE SELECTION:

Key characteristics that had substantial associations with the outcome variable, including age, BMI, insulin, and glucose, were chosen for model training based on the insights gained from data visualization.

7.6 MODEL DEVELOPMENT:

A Support Vector Classifier (SVC) with a linear kernel was used to create the model. The dataset was divided into subgroups for testing (20%) and training (80%). The testing data was used to validate the model after it had been trained on the training data.

7.7 MODEL EVALUATION:

Metrics including accuracy, precision, recall, confusion matrix, and F1-score were used to evaluate the model's capacity to reliably predict diabetes.

Fig. 7.7 Heatmap of confused matrix

7.8 MODEL DEPLOYMENT:

Finally, the trained model was deployed using **Flask**, creating a web application that allows users to input their data and receive real-time predictions about their diabetes risk.

Fig. 7.8 Model Deployment

8. CONCLUSIONS:

This study effectively applies machine learning to predict diabetes using the Pima Indians Diabetes dataset. A Support Vector Classifier (SVC) with a linear kernel was used, yielding an accuracy of about 73%. The initiative emphasizes the importance of features such as glucose, insulin, BMI, and age in predicting diabetes, particularly their involvement in early identification. A systematic technique was used to preprocess the dataset to resolve missing values and scale it to improve model performance. The successful implementation of a web application using Flask proves the model's practical applicability by offering a user-friendly platform for real-time diabetes risk assessment.

9. FUTURE SCOPE:

- Model Improvement: Utilize advanced machine learning models (e.g., Random Forest, XGBoost, or neural networks) and optimize hyperparameters for better accuracy.
- Data Expansion: Incorporate larger, more diverse datasets and address class imbalance using techniques like SMOTE.
- Feature Exploration: Add relevant features like genetics, lifestyle, and family history to enhance predictions and apply dimensionality reduction methods for efficiency.
- Integration: Deploy the model in clinical settings and integrate with EHR systems for real-time decision support.
- Technological Advancement: Expand the application to mobile or cloud platforms and integrate wearable device data for continuous monitoring.
- Extended Applications: Adapt the model for other chronic diseases like cardiovascular conditions or kidney disease.
- Collaborations: Work with healthcare professionals to validate and refine the model for practical use.

10. REFRENCES:

1. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Pima Indians Diabetes Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/diabetes.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
3. Brownlee, J. (2016). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. Machine Learning Mastery.
4. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
5. Flask Documentation. Flask: A Web Framework for Python. Retrieved from https://flask.palletsprojects.com/.
6. Chris Albon. (n.d.). Machine Learning with Python Tutorials. Retrieved from https://chrisalbon.com/.
7. Seaborn Documentation. (n.d.). Python Visualization Library. Retrieved from https://seaborn.pydata.org/.
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.

9. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.