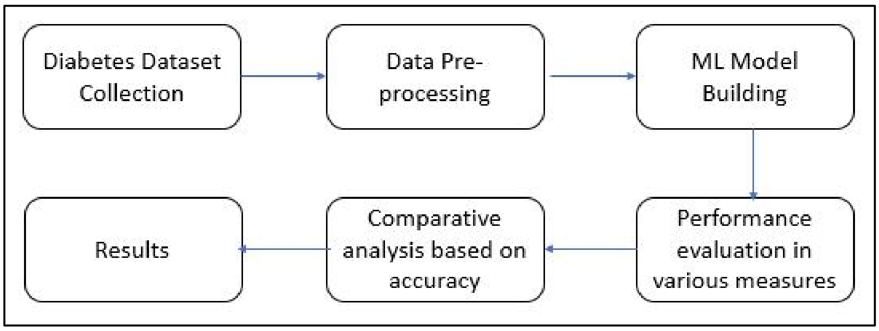
RESEARCH PAPER SUMMARY


Diabetes Dataset Collection
Data Pre-processing
ML Model Building
Performance evaluation in various measures
Comparative analysis based on accuracy
Results

ABSTRACT

Machine learning algorithms have been effectively used in many areas, including health care sectors as well. It is used to extract useful information from the database and gives a valid predicted outcome. The research paper analyses different ML algorithms to build the best model for diabetes prediction.

SUMMARY

Diabetes prediction model uses machine learning algorithms for improving remote healthcare monitoring systems for diabetic patients. PIMA Indian Diabetes Dataset [PIDD] plays a major role in research in diabetes. Most of the authors have developed prediction model for diabetes using PIDD because of its simplicity and uniqueness. Data – processing consists of three methods namely data exploration, data cleaning and model selection with evaluation phase.

The first stage helps in analyzing and exploring the dataset to have a better understanding of it. In data cleaning stage, we remove null points and outliers. Feature engineering techniques have been employed, which recursively removes unimportant attributes by which accuracy increases. We can see that by providing four features to the model we obtain the best accuracy score i.e., Pregnancies, Glucose, BMI, Diabetes Pedigree Function.

The model is trained using various algorithms such as KNN, SVM and Naive Bayes, random forest, logistic regression, decision tree by splitting the dataset to 80% training data and 20% test data.

Various evaluation metrics such as Accuracy, ROC curve, Precision, Recall, Confusion matrix, Cross validation are used to find the best model for diabetes prediction.

RESULT

Machine learning techniques were used for analyzing and predicting diabetes. According to the research paper referred, algorithms such as KNN, SVM, logistic regression, random forest, decision tree, naive bayes and gradient boost were used. From the simulation results, which was obtained, it has been proved that the Random Forest provided better accuracy of 98%. It is also stated that feature engineering, feature selection, outliers' removal and other data cleaning and data mining techniques will provide better accuracy when compared with simple prediction using simple machine learning algorithms

METHODOLOGY

ALGORITHMS USED

- K Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Naive Bayes

MODEL EVALUATION TECHNIQUES USED

- Confusion matrix
- Classification report
- K fold cross validation technique
- ROC Curve with Area under curve (AUC)
- Learning Curve
- Grid search method

EXPLORATORY DATA ANALYSIS AND VISUALISATION

Descriptive Statistics – It calculates summary statistics such as mean, median etc. to understand the distribution. It also gives the shape of the dataset and the respective data type of each feature/column.

Data Visualization – Few plots are used for understanding and visualizing the diabetes dataset. Count plot deals with the balance of the dataset. Histogram is to check if the data is normally distributed or skewed. Box plot is used for detection of outliers. Scatterplot helps to understand the relationship between features. Finally, a heatmap is used for measuring the strength of association between two features which indicates the intensity of correlation between them.

Missing Data Analysis – Identifying and handling missing values in the dataset is very crucial in data exploration.

Feature Distribution – By analyzing the distribution of features, it helps in prioritizing the features for model development and feature selection. A bar graph plots the top 5 important features of the dataset by performing feature scaling followed by feature selection and fitting it in the logistic regression model.

K NEAREST NEIGHBOR (KNN)

K value is taken to be 10, for more accuracy. It is also specified to use “Euclidean” distance measure explicitly.

The predicted accuracy is 78 % and the K fold cross validation accuracy for k value of 10 is 80 %.

We also plotted a graph for no. of nearest neighbors ranging from 1 to 19 versus the corresponding accuracy.

Following another plot briefs about K fold cross validation versus its accuracy for various K values. Wherein the K value for K fold cross validation is 12 in count ranging from 5 to 50 and the K value for KNN are chosen to be 2,3,4,5,6.

A plot of ROC curve with an AUC of 86 % and 10-fold cross validation accuracy of 80 % suggests that the model has good discriminative power and exhibits robustness across multiple iterations.

The learning curve for KNN, which relates the no. of training samples with accuracy for training accuracy and validation accuracy. We can see that as the training samples increase, the accuracy also increases.

SUPPORT VECTOR MACHINE (SVM)

A pair plot is plotted to analyze if the dataset is a linear or a nonlinear one. From the plot it is evident that the data is non – linearly distributed, so we can use RBF or polynomial kernels to proceed.

We compare the performance metrics and accuracy of both RBF and polynomial kernels to choose the best one among them. After the analysis RBF was found to be more efficient with an accuracy of 81 %.

By the AUC value, we can see that the overall performance of a classifier is 88%, which is slightly greater than KNN.

The learning curve of SVM model shows that the training and test converges together. Thus, we conclude that the performance of the SVM model tends to improve much with a larger number of training samples.

NAIVE BAYES

We evaluate the accuracy and performance metrics for Naïve Bayes, wherein the accuracy obtained equals 76% and its respective 10-fold cross validation is 75 %. From this we can see that there is not much deviation from the accuracy obtained and the cross-validation accuracy.

A plot for the ROC curve is plotted with area under curve accuracy as 82 %.

In the learning curve for Naive Bayes, we see that the training accuracy and validation accuracy almost collides indicating that both increases with increase in the number of trainings samples.

MODEL COMPARISON

We compared the 3 ML models using grid search method. The accuracy obtained from Grid Search represents the performance of the model with the optimized hyperparameters on the validation set.

It is expected to be higher compared to the accuracy obtained from cross-validation because Grid Search selects the hyperparameters that perform best on the validation set. KNN and SVM record an accuracy of 88% and 90% respectively.

We also compared the cross-validation accuracies of the three ML models using a bar graph for better visualization.

We have analyzed the dataset thoroughly using various techniques and effectively visualized the results using seaborn and mat plot libraries in the pre-processing stage.

RESULTS OBTAINED

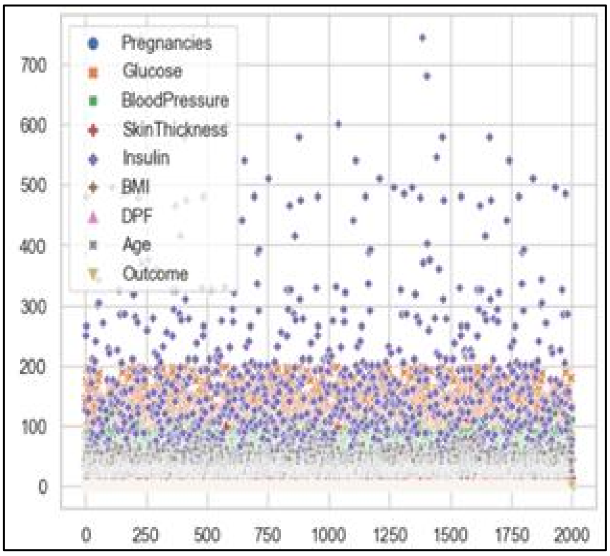
EVLAUATION METRICS SCORES

ALGORITHM	ACCURACY	AUC	PRECISION	RECALL	GRID SEARCH
KNN	78 %	0.86	76%	71%	88%
SVM	81 %	0.88	80%	75%	90%
NAIVE BAYES	76 %	0.82	72%	71%	-

ACCURACY SCORES IN K FOLD CROSS VALIDATION

ALGORITHM	ACCURACY
K NEAREST NEIGHBOUR	80%
SUPPORT VECTOR MACHINE	76%
NAIVE BAYES	75%

PLOTS



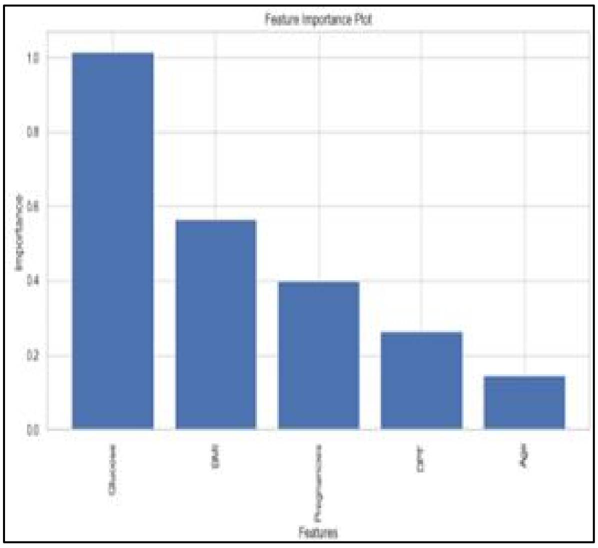
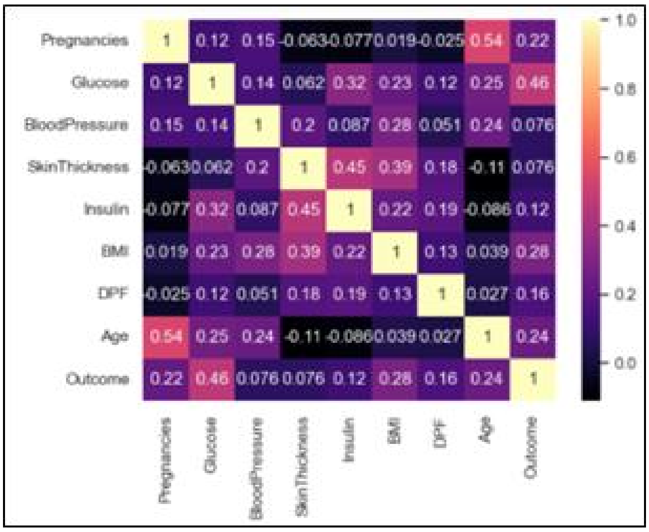


Fig 1.Scatterplot For the Dataset
Shows the relationship between each feature

Fig 2.Feature Importance Plot
The most important features of the dataset



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
Pregnancies	1	0.12	0.15	-0.063	-0.077	0.019	-0.025	0.54	0.22
Glucose	0.12	1	0.14	0.062	0.32	0.23	0.12	0.25	0.46
BloodPressure	0.15	0.14	1	0.2	0.087	0.28	0.051	0.24	0.076
SkinThickness	-0.063	0.062	0.2	1	0.45	0.39	0.18	-0.11	0.076
Insulin	-0.077	0.32	0.087	0.45	1	0.22	0.19	-0.086	0.12
BMI	0.019	0.23	0.28	0.39	0.22	1	0.13	0.039	0.28
DPF	-0.025	0.12	0.051	0.18	0.19	0.13	1	0.027	0.16
Age	0.54	0.25	0.24	-0.11	-0.086	0.039	0.027	1	0.24
Outcome	0.22	0.46	0.076	0.076	0.12	0.28	0.16	0.24	1

Fig 3. Heatmap For the Dataset

The heatmap plotted shows the correlation between all the features. According to the scale given, lighter the colour, more correlation the two features.

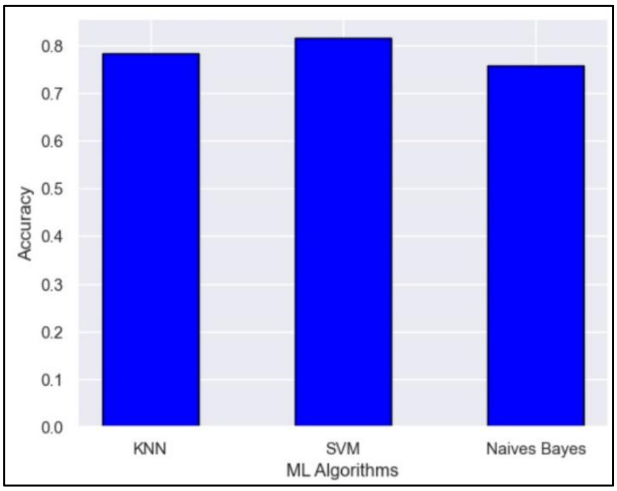
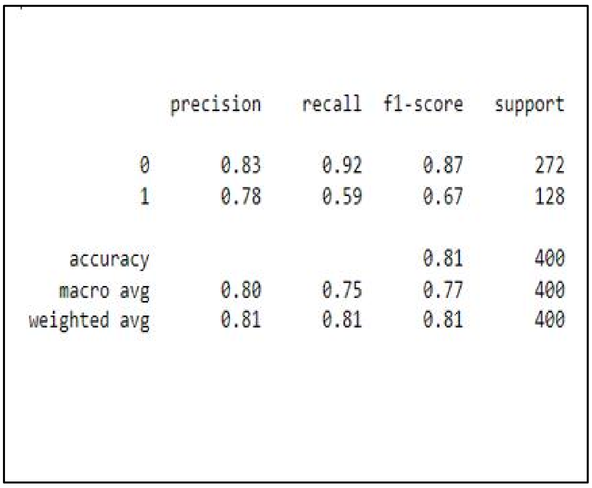


Fig 4. Algorithms Vs Accuracy



```
              precision    recall  f1-score   support

           0       0.83      0.92      0.87       272
           1       0.78      0.59      0.67       128

    accuracy                           0.81       400
   macro avg       0.80      0.75      0.77       400
weighted avg       0.81      0.81      0.81       400
```

Fig 5. Confusion matrix for SVM

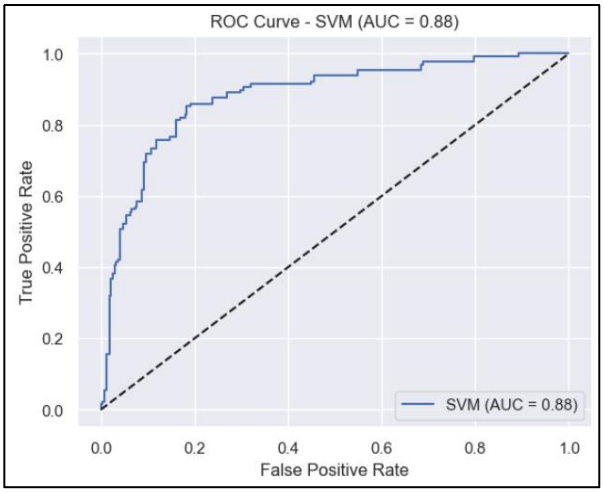


Fig 6. ROC Curve for SVM

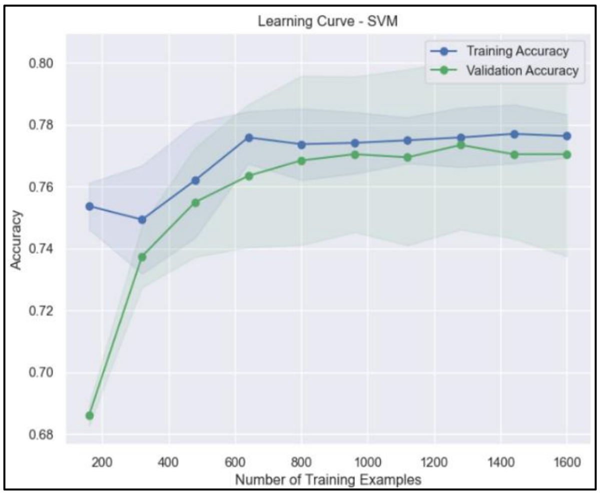


Fig 7. Learning Curve for SVM

RESULT AND DISCUSSION

RESULT

By considering multiple evaluation metrics and their values, we can see that SVM exhibits the highest calculated accuracy of 81 % and consistently performs well across various evaluation measures highest accuracy, AUC, precision, and grid search value among the three algorithms, suggesting that SVM has the potential to provide more accurate and reliable predictions than KNN and Naive Bayes. So, we chose SVM to be the most efficient algorithm for the diabetes prediction model according to the analysis done. Additionally, while KNN performs relatively well in some metrics such as cross-validation accuracy, it lags behind SVM in terms of AUC, precision, and grid search value.

INTUITION OF SVM FOR THE DATASET

SVM is known to handle high-dimensional data effectively, has the capability to handle non-linearly separable data using different kernel functions, aims to find a decision boundary that maximizes the margin between the positive and negative instances, handles imbalances in positive and negative instances by oversampling or under sampling, it also efficiently optimizes the decision boundary based on the support vectors rather than considering all training instances.

STRENGTH & WEAKNESS OF SVM

SVM is effective in high dimensional spaces, can capture nonlinear relationships between input features and target variable, aims to find a decision boundary with maximal margin which helps in achieving good generalization. The disadvantages of SVM can be such as they are sensitive to noise and outliers, memory requirements increase with increase in size of dataset, it can also be difficult to interpret the prediction as sometimes decision boundaries can be complex.

KNN, BAYES LIMITATIONS

KNN is typically a lazy learner who is not inherently equipped to handle missing data, it can be inefficient and intensive while handling large datasets. While Naïve Bayes assumes independence between features, which may not hold true in all cases, which can lead to suboptimal performance when there are strong dependencies among features.

LIMITATIONS OF THE MODEL

The applicability of the diabetes prediction model may be limited to the specific context and population represented in the PIDD dataset. Factors such as demographic characteristics, geographic location, or healthcare system variations may influence the generalizability of the model to other populations or settings.

PRACTICAL IMPLICATIONS

In the context of remote monitoring and telehealth, the diabetes prediction model can be integrated into digital health platforms. The model's predictions and outcomes can be used to generate insights into the factors influencing diabetes and the collected data can also be anonymized and aggregated to further research on diabetes risk factors, patterns, and trends. The diabetes prediction model can be used to identify individuals who are at higher risks of developing diabetes in the future.

REFERENCES

S. S et al., "A Comparative Analysis of Diabetes Prediction Models using Machine Learning Algorithms," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 261-265, doi: 10.1109/ICACCS54159.2022.9785280.

Y. Dubey, P. Wankhede, T. Borkar, A. Borkar and K. Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 60-63, doi: 10.1109/BECITHCON54710.2021