Data Mining on Diabetes Detection

Presented by

Rupak Upadhyay (202222020)

Aditya Pradhan (202222017)

Asha Kumari (202222018)

Under the Guidance of

Mr. Gaurav Pradhan, Asst. Professor, Dept. of CA, SMIT

Content

- Introduction
- General Overview of the problem
- Aim and Objective
- Feasibility Study
- Literature Study
- Problem Definition
- Proposed Solution Strategy
- Project Plan
- Workflow Chart
- Method and Methodology
- Result and Discussion
- Summary and Conclusion
- References

Introduction

- Diabetes is a long-term health problem that happens when the body has trouble handling the sugar in our blood, which we call glucose.
- There are mainly two types of diabetes: Type 1 and Type 2. There’s also gestational diabetes that can occur during pregnancy
- Type 1 diabetes is known for a deficit in insulin production and requires daily administration of insulin.
- Type 2 diabetes is a condition that happens because of a problem in the way the body regulates and uses sugar as a fuel.

General Overview of the Problem

- In the current times, many medical diagnoses are performed by machine learning, as they are accurate, and fast.
- Most of the algorithms present are highly optimized and complex for a regular person with no coding knowledge to understand.
- This research aims to search for an efficient way of detecting diabetes i.e., getting good accuracy while keeping the complexity of the used algorithms as low as possible

Aim and Objective

- The majority of the machine learning models currently available are complex.
- So to get a high success rate in detecting diabetes the developers of these models optimize the algorithms as much as possible.
- The challenge of this research lies in the fact that to get higher accuracy a more complex model is required.
- The aim and objective of this study is to make the model as simple as possible.

Feasibility Study

- **Economic Feasibility**
  - The tools used in this study will be open-source.
  - So, there will be no need for any external funding.
- **Technical Feasibility**
  - This study will use a modern laptop with a reasonable processing capability.
  - This study will use Python libraries (sci-kit-learn, pandas, numpy) and IDE (Visual Studio Code).
- **Operational Feasibility**
  - The team conducting this study consists of MCA students and is equipped with the knowledge and skills required to complete this study.
- **Schedule Feasibility**
  - Will be completed within the stipulated time.

Literature Survey

Table. Literature Survey

Serial No.	Name of article	Detection Type	Findings	Research Gap	Relevant to Proposed work
1.	Luís Chaves et. al. [2021] "Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study"[2] 2021 MDPI	Supervised	**Methodology:** Classification: Naïve Bayes, Neural Networks, AdaBoost, kNN, Random Forest, Support Vector Machine **Dataset Used:** The data was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh **Results:** Neural Network achieved 98.1% accuracy	The data set has limitations since it does not consider family history of diabetes, consumption of certain prescription drugs, smoking, and sleep deprivation.	Classification: Random Forest, Naïve Bayes, kNN, SVM

Literature Survey

Contd..

Table. Literature Survey (contd..)

Serial No.	Name of article	Detection Type	Findings	Research Gap	Relevant to Proposed work
2.	S. Saru et. al. [2019] “Analysis and Prediction of Diabetes Using Machine Learning”[1] 2016 SSRN	Supervised	**Methodology:** Feature Selection: Principal Component Analysis Classification: Random Forest, k-Nearest Neighbors, Decision Tree, Support Vector Machine, Logistic Regres sion **Dataset Used:** PIMA Indian Diabetes Data **Results:** Decision Tree achieved 87% accuracy	Small amount of sample data was used. Only a single data set used, in future multiple data set can be used for detection.	Feature Selection: Principal Component Analysis Classification: Random Forest, Support Vector Machine, k-Nearest Neighbors

Literature Survey

Contd..

Table. Literature Survey (contd..)

Serial No.	Name of article	Detection Type	Findings	Research Gap	Relevant to Proposed work
3.	F. A. Khan et. al. [2021] “Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review”[3] 2021 IEEE	Supervised	**Methodology:** Classification: Artificial Neural Networks, Support Vector Regression, Decision Tree, Support Vector Machine, Navie Bayes, Random Forest, Instance Based Learner Classifier, Quadratic Discriminant Analysis **Dataset Used:** Dataset collected from an online portal and a college medical hospital **Results:** ANN achieved 89% accuracy	It is evident from the analysis of the schemes in all classes that most of them suffer from either single data input parameter or the parameter selection process is not optimal.	Classification: Artificial Neural Networks, SVM, Random Forest, Naïve Bayes

Literature Survey

Contd..

Table. Literature Survey (contd..)

Serial No.	Name of article	Detection Type	Findings	Research Gap	Relevant to Proposed work
4.	Wee et. al. [2023] "Diabetes Detection Based on Machine Learning and Deep Learning Approaches"[7] 2023 Springer	Supervised	**Methodology:** Feature Selection: Pearson Correlation Coefficient, Fuzzy Support Vector Machine, F-Score Feature Selection Classification: Support Vector Machine, Random Forest **Dataset Used:** Pima Indians Diabetes Data **Results:** Random Forest achieved 79.26% accuracy	Standard procedure is not proposed for collecting the required data for further investigation and improving the classification models. Features such as ECG are not proven to have a direct relationship to diabetes	Feature Selection: F-Score Feature Selection Classification: Random Forest, SVM

Problem Definition

- The majority of the machine learning models currently available are complex.
- So to get high success rate in detecting diabetes the developers of these models optimize the algorithms as much as possible.
- The challenge of this research lies in the fact that to get higher accuracy a more complex model is required.

Proposed Solution Strategy

- To get high accuracy while keeping the complexity of the machine learning model low, the team will use five feature selection algorithms, to select features from a dataset.
- Then use six classification algorithms to detect diabetes, from these newly acquired datasets of selected features. The plan is to keep the parameters of these algorithms as simple as possible.
- The steps:
  - Data Acquisition and Preprocessing
  - Feature Selection
  - Classification
  - Evaluation

Project Plan

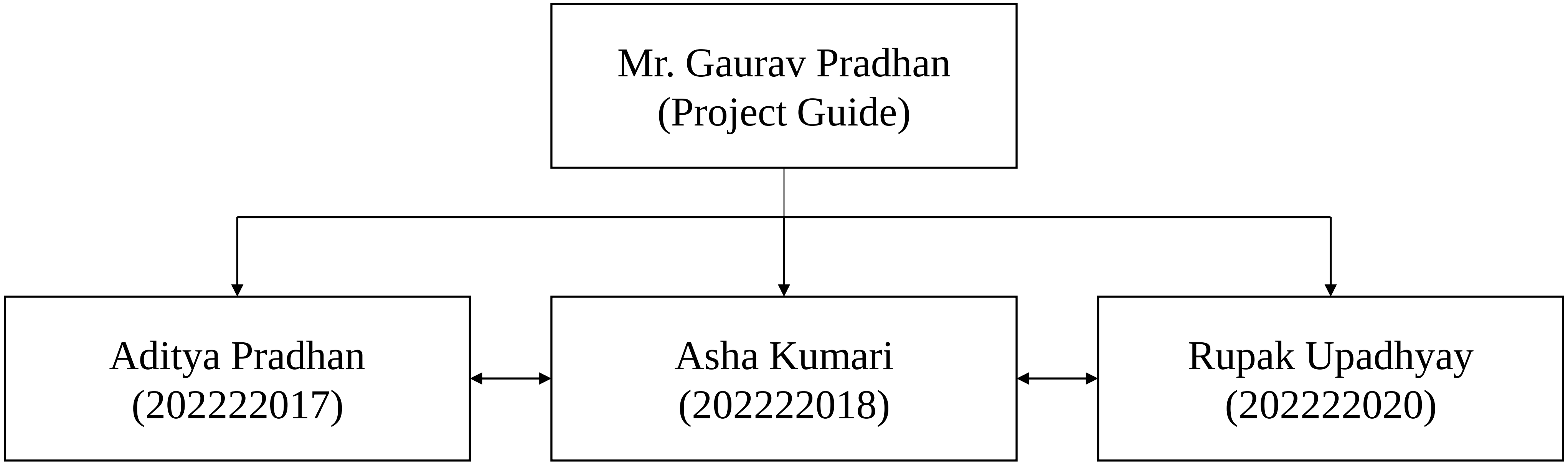

Fig. Team Structure

Project Plan

Contd..

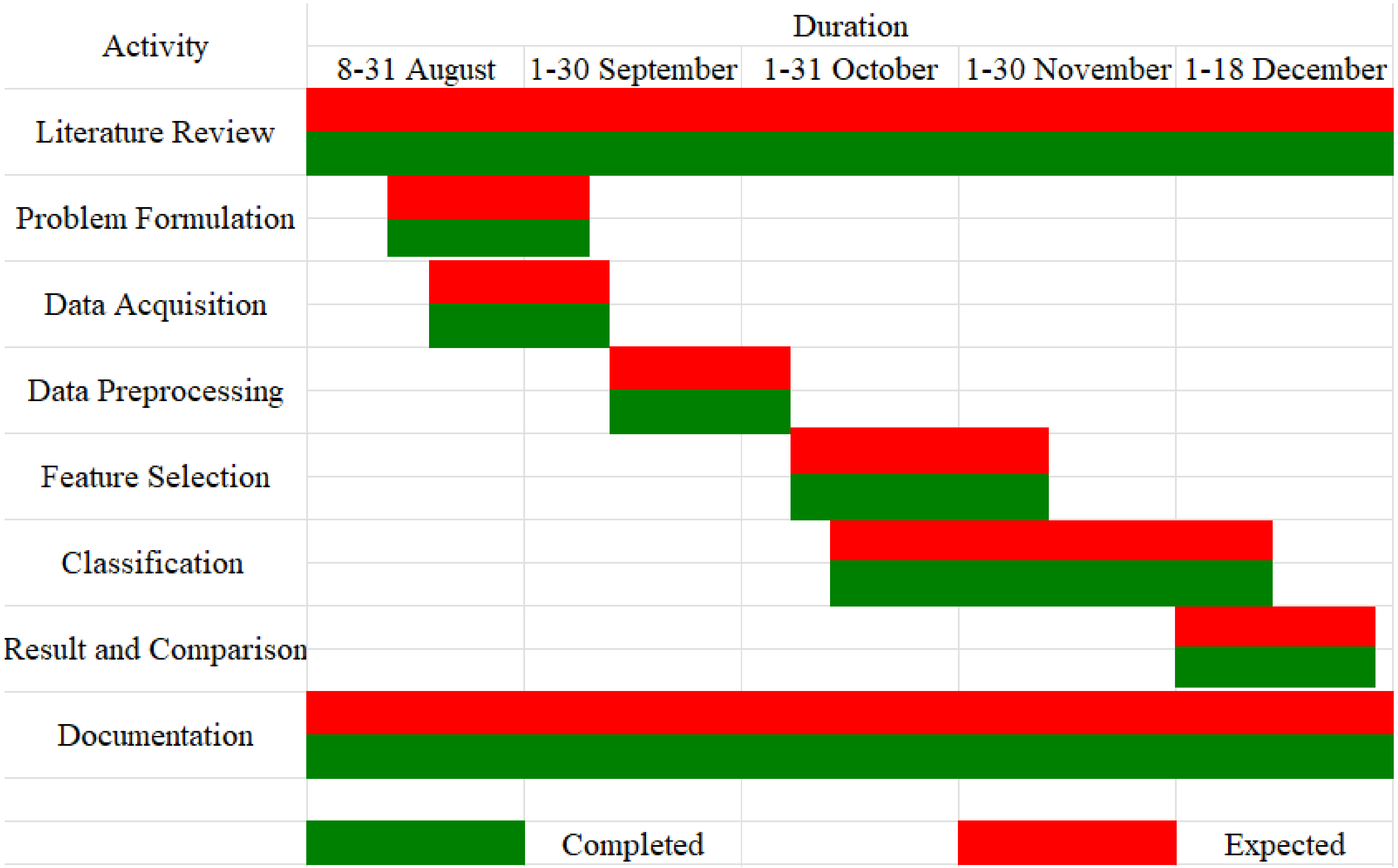

Fig. Gantt Chart

Project Plan

Contd..

Hardware Requirement

Processor: Intel Core i3 7th Gen+ or AMD Ryzen 3 2nd Gen+

Memory: 4GB DDR3 RAM

Storage: 50GB

Software Requirement

Operating System: Microsoft Windows 7 and above

Programming Language: Python

Development Environment: Visual Studio Code, Weka

Workflow Chart

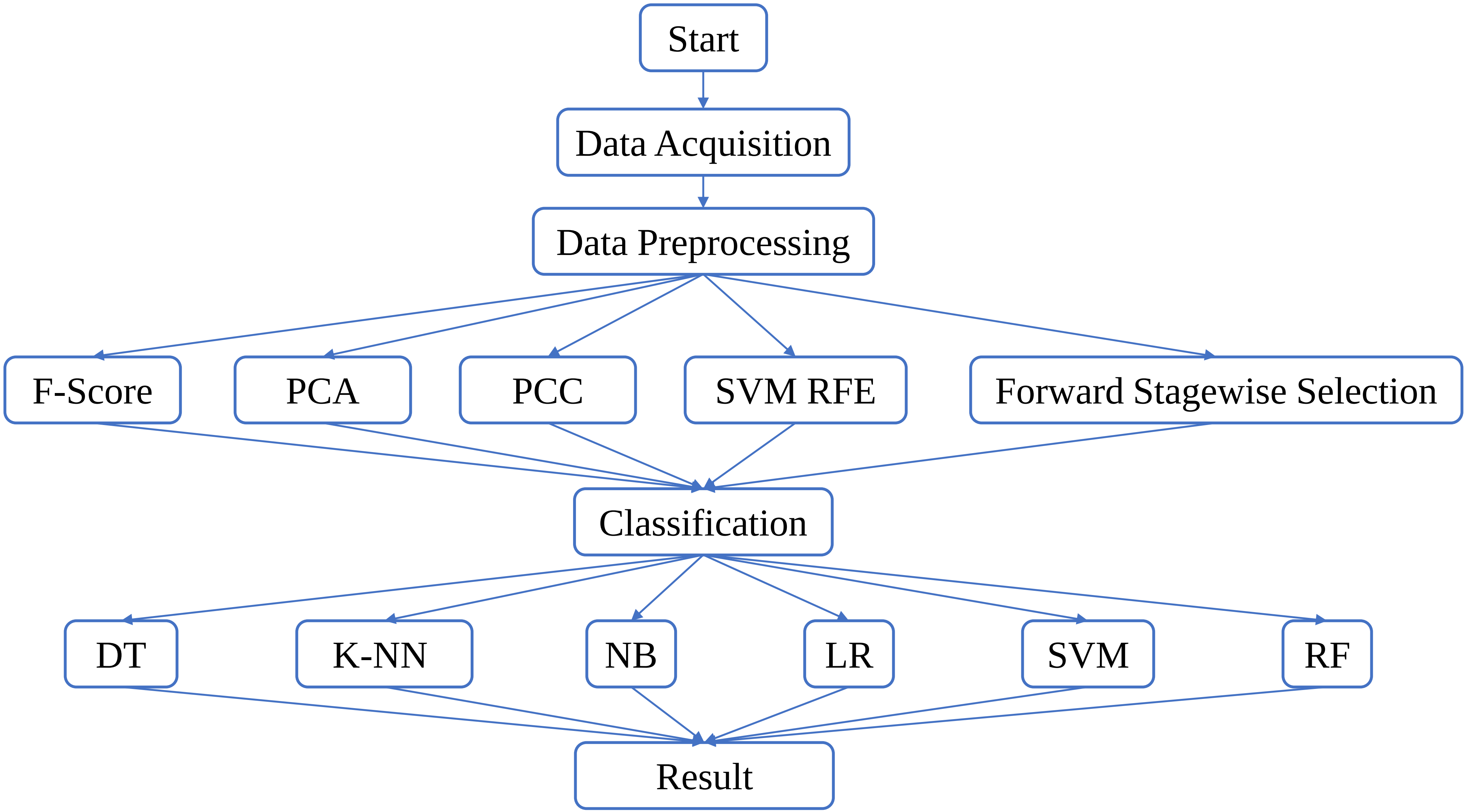

Fig. Workflow Chart

Method and Methodology

- **Data Acquisition**
  - Pima Indians Diabetes Database 768 individual records with eight features
  - Dataset2: 2535 individual records with 19 features
- **Data Preprocessing**
  - Weighted k-NN
  - KNNImputer(n_neighbors = 10, weights = ‘distance’)

Method and Methodology

- **Feature Selection**
  - F-Score:
  - Selector = SelectKBest(score_func = f_classif, k = 5)
  - Principal Component Analysis:
  - PCA(n_components = 5)
  - X_pca = pca.fit_transform(X_scaled)

Method and Methodology

Contd..

- **Feature Selection**
  - Pearson Correlation Coefficient:
  - correlation_matrix = df.corr()
  - corelation_with_target = correlation_matrix[‘tar_var’].abs()
  - sorted_features = correlation_with_target.sort_value(ascending = False)
  - sorted_features[1:6]

Method and Methodology

- **Feature Selection**
  - Support Vector Machine Recursive Feature Elimination:
  - svm = SVC(kernel = ‘linear’)
  - rfe = RFE(estimator = svm, n_features_to_select = 5)
  - rfe.fit(X_train, y_train)

Method and Methodology

Contd..

- **Classification**
  - Decision Tree: weka.classifiers.trees.J48 -C 0.25 -M 2
  - K-Nearest Neighbor: weka.classifiers.lazy.IBk -K 1 -W 0 -A “weka.core.neighboursearch.LinearNNSearch -A \ “weka.core.EuclideanDistance -R first-last \””
  - Logistic Regression: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4

Method and Methodology

- **Classification**
  - Naïve Bayes: weka.classifiers.bayes.NaiveBayes
  - Random Forest: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
  - Support Vector Machine: weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Result and Discussion

Table. Description ton of PIMA Indian Diabetes Database[19]

Serial No.	Features	Description
1	Pregnancies	Number of pregnancies
2	Glucose	Glucose tolerance test using plasma glucose concentration
3	BloodPressure	Diastolic blood pressure(mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin (mu U.ml)
6	BMI	Body mass index
7	DiabetesPedigreeFunction	Diabetes pedigree Function
8	Age	Age in years
9	Outcome	Class label(0 or 1)

Result and Discussion

Table. Description ton of Dataset2

Serial No.	Features	Description
1	Sex	The gender of the patient
2	Age	Age of the patient in years
3	History of high blood pressure	High blood pressure history of the patient (Yes or No)
4	History of use of drugs for high blood pressure	Use of drugs for high blood pressure (Yes or No)
5	Systolic blood pressure	Systolic blood pressure in mm of Hg
6	Diastolic blood pressure	Diastolic blood pressure in mm Hg
7	Height	Height of the patient in cm
8	Weight	Weight of the patient in kg
9	BMI	Body mass index
10	history of diabetes	History of diabetes

Result and Discussion

Contd..

Table. Description ton of Dataset2

Serial No.	Features	Description
11	Family history of diabetes	Diabetes in the patient's family
12	History of aborted baby	Abortion performed on the patient
13	History of gestational diabetes	Gestational diabetes in the patient (Yes or No)
14	History of pregnancy	Pregnancy history of the patient (Yes or No)
15	FBS	Fasting Blood Sugar test (mg/dL)
16	Cholesterol	Cholesterol level measured in mg/dL
17	HDL	High-density lipoprotein measured in mg/dL
18	Triglyceride	Triglycerides level measured in mmol/L
19	result of high blood pressure screening	Class label (Positive or Negative or Old patient)
20	result of diabetes screening	Class label (Positive or Negative or Old patient)

Result and Discussion

Contd..

Table. Features selected using F-score

Pima Indians Dataset	Dataset2
• Glucose • SkinThickness • Insulin • BMI • Age	• Systolic blood pressure • Diastolic blood pressure • BMI • Family history of diabetes • History of aborted baby

Result and Discussion

Table. Features selected using PCC

Pima Indians Dataset	Dataset2
• Glucose • SkinThickness • Insulin • BMI • Pregnancies	• Systolic blood pressure • Diastolic blood pressure • BMI • Family history of diabetes • result of high blood pressure screening

Result and Discussion

Contd..

Table. Features selected using SVM RFE

Pima Indians Dataset	Dataset2
• Glucose • DiabetesPedigreeFunction • Age • BMI • Pregnancies	• History of diabetes • History of pregnancy • result of high blood pressure screening • Family history of diabetes • History of aborted baby

Result and Discussion

Contd..

Table. Features selected using Forward Stagewise Selection

Pima Indians Dataset	Dataset2
• Glucose • Insulin • Age • BMI • DiabetesPedigreeFunction	• Age • BMI • HDL • Family history of diabetes • History of aborted baby

Result and Discussion

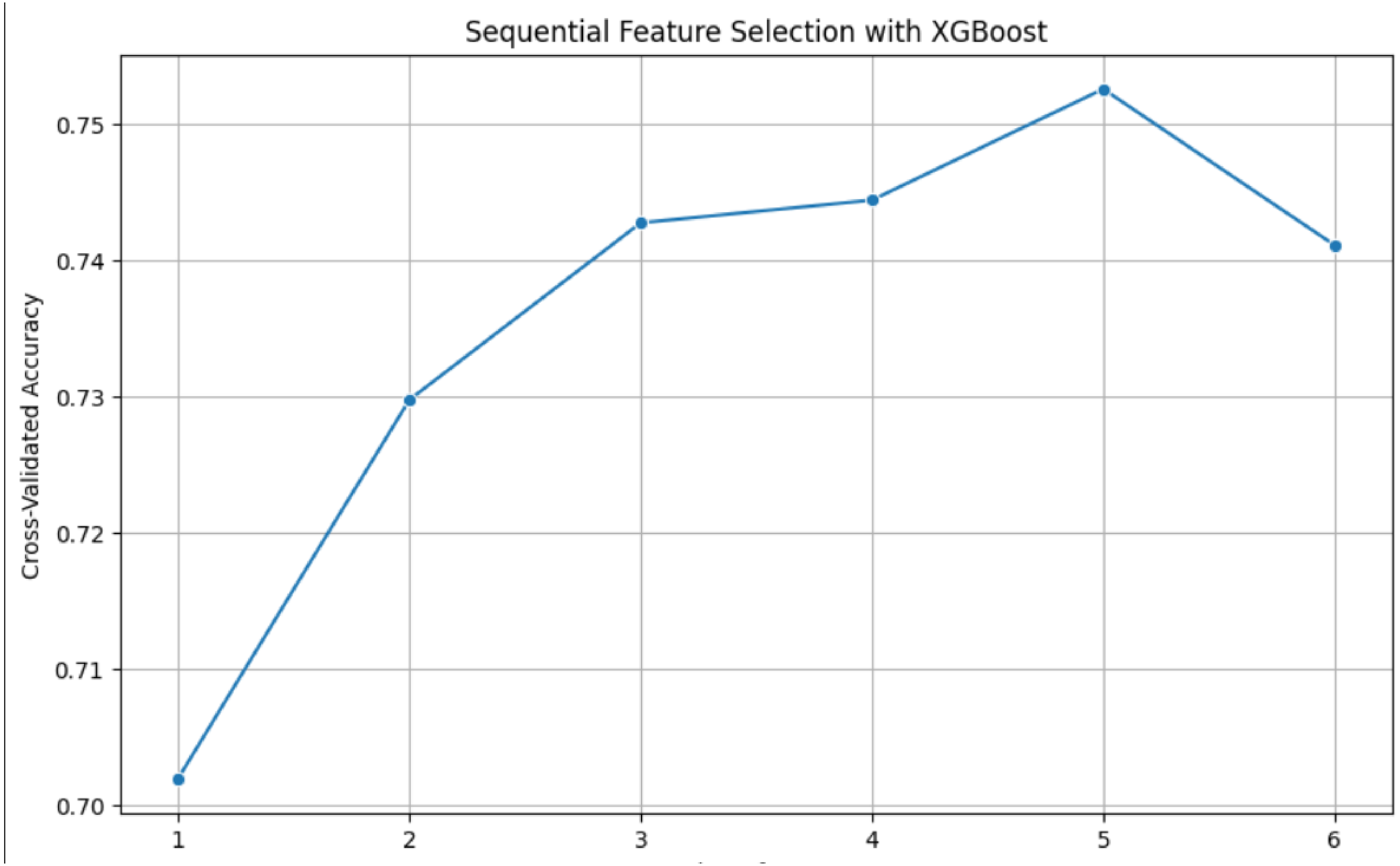

Fig. Graph illustrating the performance of Forward Stagewise selection

Result and Discussion

Table. Result of Classification algorithm on PINA Indians Dataset

Feature Selection / Classification	Forward Stagewise Selection	F-score	PCA	PCC	SVM_RFE
DT	65.1042	65.1042	65.1042	65.1042	65.1042
kNN	67.0573	64.4531	65.1042	65.8854	63.4115
LR	57.1615	64.9740	65.1042	65.4948	59.7656
NB	67.3177	69.5313	35.4167	69.2708	68.6198
RF	64.7135	65.8854	65.1042	66.6667	66.4063
SVM	65.2344	65.3646	65.1042	68.6198	64.8438

Result and Discussion

Contd..

Table. Result of Classification algorithm on Dataset2

Feature Selection / Classification	Forward Stagewise Selection	F-score	PCA	PCC	SVM_RFE
DT	91.4004	91.4004	91.4004	91.4004	91.4004
kNN	91.4398	91.0848	91.3609	90.6114	91.4004
LR	88.9152	88.7968	91.4004	87.9684	91.4004
NB	91.5187	91.3215	8.6391	91.1637	91.4004
RF	90.6509	91.4793	91.4004	91.5976	91.4004
SVM	91.5582	91.5187	91.4004	91.4398	91.4004

Result and Discussion

Contd..

Table. Comparison of Accuracy

Study \ Algorithm	DT	kNN	LR	NB	RF	SVM
[25]	81.3000	-	83.8600	-	87.4000	86.0200
[27]	-	83.3300	-	-	-	-
[1]	-	97.3100	-	86.9200	96.9200	97.1200
[11]	94.4400	93.7900	-	79.8400	-	-
[9]	-	78.4000	-	92.7000	96.9000	90.8000
This study	91.4004	91.4398	91.4004	91.5187	91.4793	91.5582

Result and Discussion

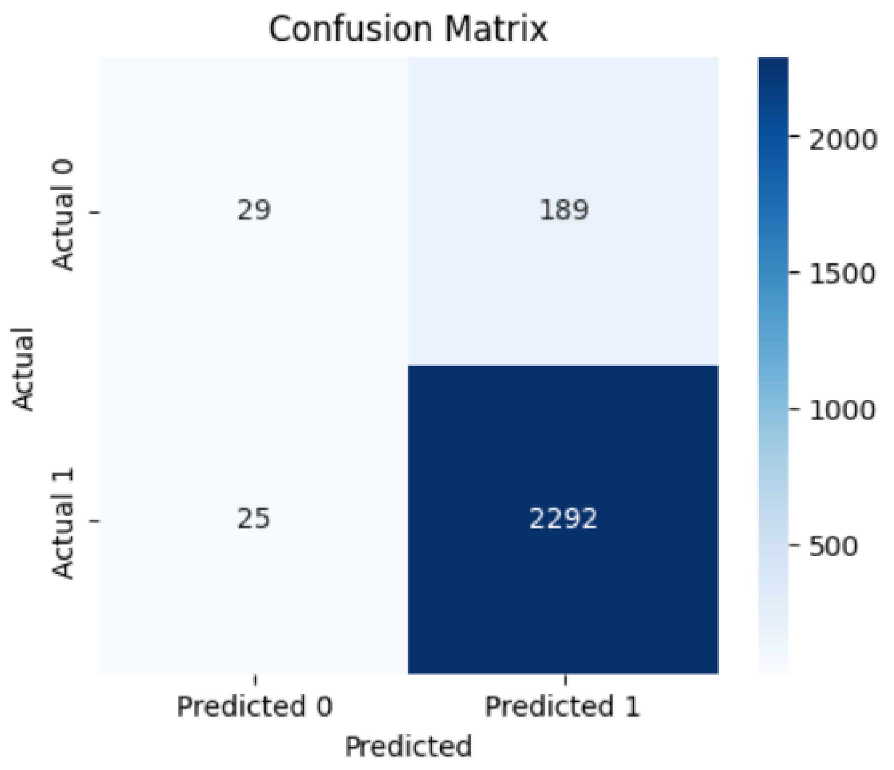

Fig. Confusion Matrix visualizing the result of SVM classifier

Summary and Conclusion

- **Progress Summary**
  - Outlined the challenges associated with existing learning models or detecting diabetes, wherein there is a trade-off between accuracy and complexity.
  - This study also proposes a well-structured solution strategy that aligns with the identified challenges.
  - While this study was in progress the study team reviewed many research papers to keep themselves up to date with the current trends and best practices in the field
  - This study is on schedule and will be completed by the established deadline.

Summary and Conclusion

- **Conclusion**
  - This study aimed to address the challenges of obtaining high accuracy in detecting diabetes while keeping the complexity of the machine learning model low
  - Two datasets were collected, namely Pima Indians Diabetes Database[19] and Datase2 which were pre-processed using a weighted kNN approach.
  - The features that were obtained showed that different algorithms selected different sets of features.
  - The highest accuracy achieved was 91.5582 using the SVM with features obtained from the forward stagewise selection.

Summary and Conclusion

- **Future Scope**
  - This study only considers a handful of feature selection and classification algorithms.
  - Further going on more classification algorithms can be used with their default parameters to see if they give more accurate results.

References

[1] L. Chaves and G. Marques, “Data mining techniques for early diagnosis of diabetes: A comparative study,” *Appl. Sci.*, vol. 11, no. 5, pp. 1–12, Mar. 2021, doi: 10.3390/app11052218.

[2] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-16407-5.

[3] “Diabetes.” Accessed: Nov. 17, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] “Diabetes Facets and Figures | International Diabetes Federation.” Accessed: Nov. 17, 2023. [Online]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/

[5] N. Jothi, N. A. Rashid, and W. Husain, “Data Mining in Healthcare - A Review,” *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015, doi: 10.1016/j.procs.2015.12.145.

[6] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, “Mining association rules in graphs based on frequent cohesive itemsets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[7] S. Saru, “ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING,” 2019. [Online]. Available: https://ssrn.com/abstract=3368308

[8] U. Ahmed *et al.*, “Prediction of Diabetes Empowered With Fused Machine Learning,” *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

References

[9] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, “Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm),” *Diagnostics*, vol. 11, no. 9, Sep. 2021, doi: 10.3390/diagnostics11091714.

[10] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, “Prediction and diagnosis of future diabetes risk: a machine learning approach,” *SN Appl. Sci.*, vol. 1, no. 9, Sep. 2019, doi: 10.1007/s42452-019-1117-9.

[11] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, “Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: A comparison of four data mining approaches,” *BMC Bioinformatics*, vol. 21, no. 1, Aug. 2020, doi: 10.1186/s12859-020-03719-8.

[12] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, “A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus,” *Appl. Intell.*, vol. 52, no. 3, pp. 2411–2422, Feb. 2022, doi: 10.1007/s10489-021-02533-w.

[13] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0175-6.

[14] R. B. Lukmanto, Suharjito, A. Nugroho, and H. Akbar, “Early detection of diabetes mellitus using feature selection and fuzzy support vector machine,” in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 46–54. doi: 10.1016/j.procs.2019.08.140.

[15] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-68771-z.

[16] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

References

[17] R. Ahuja, S. C. Sharma, and M. Ali, “A diabetic disease prediction model based on classification algorithms,” *Ann. Emerg. Technol. Comput.*, vol. 3, no. 3, pp. 44–52, Jul. 2019, doi: 10.33166/AETiC.2019.03.005.

[18] A. Iyer, J. S, and R. Sumbaly, “Diagnosis of Diabetes Using Classification Mining Techniques,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 01–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101.

[19] “Pima Indians Diabetes Database.” Accessed: Nov. 18, 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[20] “Feature Selection - F-score | Kaggle.” Accessed: Nov. 18, 2023. [Online]. Available: https://www.kaggle.com/code/tanmayunhale/feature-selection-f-score

[21] “Reduce Data Dimensionality using PCA - Python - GeeksforGeeks.” Accessed: Nov. 19, 2023. [Online]. Available: https://www.geeksforgeeks.org/reduce-data-dimentionality-using-pca-python/

[22] “Feature Selection - Pearson Correlation | Kaggle.” Accessed: Nov. 19, 2023. [Online]. Available: https://www.kaggle.com/code/tanmayunhale/feature-selection-pearson-correlation

[23] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, “SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–18, 2018, doi: 10.1186/s12859-018-2451-4.

[24] “9.6 Feature Selection via Boosting.” Accessed: Nov. 19, 2023. [Online]. Available: https://jermwatt.github.io/machine_learning_refined/notes/9_Feature_engineer_select/9_6_Boosting.html

References

Contd..

[25] M. J. Uddin *et al.*, “A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh,” *Inf.*, vol. 14, no. 7, Jul. 2023, doi: 10.3390/info14070376.

[26] J. Li *et al.*, “Feature selection: A data perspective,” *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.

[27] A. Anggrawan and M. Mayadi, “Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 405–418, 2023, doi: 10.30812/matrik.v22i2.2777.

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along the completion of our project work. Whatever we have done is only due to such guidance and assistance and we would not forget to thank them. We respect and thank to **Mr. Gaurav Pradhan, Asst. Professor, Department of Computer Application,** for giving us an opportunity to do the project work in Diabetes Detection using Data Mining providing us all support and guidance which made us to complete the project on time. We are extremely grateful to him for providing such a nice support and guidance.

We are also grateful to **Dr. Samarjeet Borah,** Head of Department, Department of Computer Application, Sikkim Manipal Institute of Technology, Majitar for his guidance and suggestions during this project work.

We respect and thank **Mr. Gaurav Pradhan, Asst. Professor,** Department of Computer Application, Sikkim Manipal Institute of Technology, Majitar. for his guidance and suggestions during this project work.

We are also thankful to **Mr. Bishal Pradhan**, Mini-Project Coordinator, Department of Computer Application, Sikkim Manipal Institute of Technology for his unlisted encouragement and more over for his timely support and guidance till the completion of our project work.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of Department of computer Application which helped us in successfully completing our project work.

Aditya Pradhan (202222017)

Ahsa Kumari (202222018)

Rupak Upadhyay (202222020)

Thank You