Nicole D’Souza A020 Kunjal Shah A065 Pranav Singh A075

| Sr No | Title | Year | Software | Journal | Inference |
| --- | --- | --- | --- | --- | --- |
| 1 | Diabetes Prediction Using Artificial Neural Network | 2018 |  | International Journal of Advanced Science and Technology Vol.121 (2018) | The diabetes dataset contained 1004 samples with 9 attributes. The objective of the study was to design an artificial neural network model to determine the effective variables and their impact on diabetes. These factors were categorized as input variables and output  variables that reflect some possible levels of disease status in terms of the assessment  system. The data was entered into the JNN tool environment and then the data was trained, validated, and tested to get 87.3% accuracy. |
| 2 | Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus | 2019 |  | 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) | The authors collected a real diagnostic dataset having various attributes or risk factors of diabetes mellitus of 200 patients from a medical center. They later made a performance comparison of different ML techniques and evaluated the prediction results based on relevant risk factors. Their results showed that the C4.5 decision tree was superior for the classification of data. |
| 3 | Estimation of Gestational Diabetes Mellitus using Azure AI Services | 2019 |  | 2019 Amity International Conference on Artificial Intelligence (AICAI) | Paper presents a procedural approach for Estimation of Gestational Diabetes using Microsoft Azure AI services. Pima Indian dataset from UCI was utilized to build and cross validate the predictive model. The classification model can predict the occurrence of Gestational diabetes based on factors at early stages of pregnancy. The algorithm was examined on a dataset of 768 samples and it was able to achieve an accuracy of 77.8%. |
| 4 | Deep Learning based System Design for Diabetes Prediction | 2021 |  | 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) Pune, India. | Used a deep learning method to create a massive diabetes forecasting information processing system. The aim was to decrease the amount of false positives and negatives as much as possible in order to improve recall and precision. For prediction of diabetes ELM classifier was applied due to their quick capability of learning. With a 98.07 percent accuracy rate, DL was the most efficient and promising of the four suggested classifiers for analyzing diabetes. |
| 5 | Risk prediction of type II diabetes based on random forest model | 2017 |  | 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bioinformatics (AEEICB17) | This study proposed a type II diabetes prediction model based on random forest which was aimed at analyzing some readily available indicators. Random forest is an ensemble classifier composed of multiple decision trees, which has the advantages of high accuracy and good robustness. Dataset from the University of Virginia were used. The authors aim to expand on other indicators to predict the risk of diabetes and also update the perspective of data mining in the future. |
| 6 | Analysis of diabetes mellitus for early prediction using optimal features selection | 2019 |  | Journal of Big Data Volume  6, Article number: 13 (2019) | The objective of this research was to make use of significant features and design a prediction algorithm using Machine learning to find the optimal classifier to give the closest result compared to clinical outcomes. The proposed method aimed to focus on selecting the attributes that fail in early detection of Diabetes Mellitus using Predictive analysis. The result showed that the decision tree and random forest algorithm had the highest specificity of 98.20% and 98.00%, respectively, which hold best for the analysis of diabetic data. Naïve Bayesian outcome states the best accuracy of 82.30%. The research also generalizes the selection of optimal features from the dataset to improve the classification accuracy. |
| 7 | Type 2 diabetes mellitus prediction model based on data mining | 2017 | Waikato Environment for Knowledge Analysis toolkit | [Informatics in Medicine Unlocked](https://www.sciencedirect.com/science/journal/23529148) Volume 10, 2018, Pages 100-107 | The authors proposed a model based on data mining techniques for predicting type 2 diabetes mellitus. The main problems that they tried to solve were to improve the accuracy of the prediction model and to make the model adaptive to more than one dataset. Based on a series of preprocessing procedures, the model is composed of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were used to compare results. The conclusion showed that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, their model ensured that the dataset quality was sufficient. To further evaluate the performance of our model, they applied it to two other diabetes datasets. Both experiments' results showed good performance. |
| 8 | Diabetes Prediction: A Deep Learning Approach | 2019 |  | I.J. Information Engineering and Electronic Business, 2019, 2, 21-27 | The authors proposed a strategy for the diagnosis of diabetes using deep neural networks by training its attributes in five-fold and ten-fold cross-validation fashion. The Pima Indian Diabetes (PID) dataset was retrieved from the UCI machine learning repository database. The results on PID dataset demonstrated that deep learning approach design an auspicious system for the prediction of diabetes with prediction accuracy of 98.35% for five-fold cross-validation which was comparatively higher than other methods which are used to predict diabetes mellitus. |
| 9 | Deep learning approach for diabetes prediction using PIMA Indian dataset | 2020 |  | Journal of Diabetes & Metabolic Disorders volume  19, pages 391– 403 (2020) | This paper aimed to implement a prediction model for the risk measurement of diabetes. Diverse classifiers were used on the PIMA dataset which proved that data mining and machine learning algorithms can reduce the risk factors and improve the outcome in terms of efficiency and accuracy. Accuracy achieved by the four classifiers (DT, ANN, NB, and DL) lied within the range 90–98%, which was considerably higher than available methods. Among the four proposed classifiers, DL was considered as the most efficient and promising for analyzing diabetes with an accuracy rate of 98.07%. In the future, the authors intend to develop a robust system in the form of an app or a website that can use the proposed DL algorithm to help healthcare specialists in the early detection of diabetes. |
| 10 | Prediction of Diabetes using Classification Algorithms | 2018 |  | Procedia Computer Science: International Conference on Computational Intelligence and Data Science (ICCIDS 2018) | The authors worked towards designing a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. The three machine learning classification algorithms used were Decision Tree, SVM and Naive Bayes. Experiments were performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performance of all the three algorithms was then evaluated on several measures like Precision, Accuracy, F-Measure, and Recall. Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% compared to other algorithms. These results were then verified by making use of Receiver Operating Characteristic (ROC) curves in a proper and systematic manner |
| 11 | Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers | 2020 |  | IEEE Access : SPECIAL SECTION ON DEEP LEARNING ALGORITHMS FOR INTERNET OF MEDICAL THINGS | In this literature, the authors have proposed a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed. The weighted ensembling of different ML models is also proposed, in this literature, to improve the prediction of diabetes where the weights are estimated from the corresponding Area Under ROC Curve (AUC) of the ML model. AUC is chosen as the performance metric, which is then maximized during hyperparameter tuning using the grid search technique. All the experiments, in this literature, were conducted under the same experimental conditions using the Pima Indian Diabetes Dataset. From all the extensive experiments, our proposed ensembling classifier is the best performing classifier with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC as 0.789, 0.934, 0.092, 66.234, and 0.950 respectively which outperforms the state-of-the-art results by 2.00 % in AUC. |
| 12 | Application of data mining methods in diabetes prediction | 2017 |  | IEEE 2nd International Conference on Image, Vision and Computing | The Authors have made use of the Pima Indian Dataset and applied 5 different data mining algorithms to conduct the early prediction of diabetes, the 5 algorithms and their corresponding accuracies are Gaussian Mixture Model which gave 815 accuracy,Artificial Neural Networks with accuracy of 89% ,ELM with an accuracy of 82% ,Logistic Regression and Support Vector Machine with the accuracy of 64% and 74% respectively.The highest accuracy was found in ANN |
| 13 | Diabetes Prediction using Machine Learning Algorithms | 2019 |  | International Conference on Recent Trends in Advanced Computing 2019,ICRTAC 2019 | The authors of this paper implemented a variety of algorithms including a few ensemble learning algorithms as well.These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier,The highest accuracy was obtained by Logistic Regression at 96% followed by LDA at 94% |
| 14 | A model for early prediction of diabetes | 2019 |  | Informatics in Medicine Unlocked-Elsevier | In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. The authors findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the Apriori method. Artificial neural network (ANN), random forest (RF) and K-means clustering techniques were implemented for the prediction of diabetes. The ANN technique provided a best accuracy of 75.7%, and may be useful to assist medical professionals with treatment decisions |
| 15 | Prediction of Diabetes Using Machine Learning Algorithms in Healthcare | 2018 |  | IEEE Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018 | The authors of this paper made use of an online tool called Enthought Canopy. Enthought Canopy offers a verified scientific and analytical Python package distribution with key important integrated tools for application development, iterative data analysis and data visualization.6 algorithms i.e. KNearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) and Random Forest (RF), with SVM and KNN providing the highest accuracy i.e. 77% |
| 16 | Diabetes Prediction Using Different Machine Learning Approaches | 2019 |  | Third International Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4 | The authors of this paper implemented a variety of algorithms including Decision tree, Support Vector Machine, Naive Bayes and ANN for accuracy authentication.with the highest accuracy achieved using Support Vector Classifier and Artificial Neural Network at 82% |
| 17 | Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques | 2019 |  | Informatics in Medicine Unlocked-Elsevier | This paper proposes a new methodology i.e. using PCA to transform the initial set of features, thereby solving the problem of correlation, which makes it difficult for the classification algorithm to find relationships among the data. The PCA application will help filter out irrelevant features, thereby lowering the training time, cost, and also increasing model performance. After performing PCA analysis, the result is then passed for unsupervised clustering using K-means because of the ability of k-means to address outliers. The K-means cluster result is cleaned and Logistic Regression is applied to build a supervised classification model for the dataset The performance accuracy of the model stands at 89%. The authors have stated that accuracy further increases in a real time dataset |
| 18 | An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier | 2021 |  | Elsevier:International Journal of Cognitive Computing in Engineering | The Pima Indians Diabetes dataset has been utilized by the authors for experimentation, which gathers details of patients with and without having diabetes. The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms viz. random forest, logistic regression, and Naive Bayes for the classification. Empirical evaluation of the proposed methodology has been conducted with state-of-the-art methodologies and base classifiers such as AdaBoost, Logistic Regression,Support Vector machine, Random forest, Naïve Bayes, Bagging, GradientBoost, XGBoost, CatBoost. by taking accuracy, precision, recall, F1-score as the evaluation criteria. The proposed ensemble approach gives the highest accuracy, precision, recall, and F1\_score value with 79.04%, 73.48%, 71.45% and 80.6% respectively on the PIMA diabetes dataset. |
| 19 | Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm | 2019 |  | Fifth Information Systems International Conference 2019-Procedia Computer Science | Authors of the paper employee the Pima Indian Diabetes Dataset along with both clustering and classification to find solutions to their problem,i.e a model that predicts if a patient has diabetes and to determine what factors trigger diabetes in patient and are they interrelated, for the classification problem algorithms like Naives Bayes, KNN and Decision Tree were used.For the clustering problem Kmeans clustering was used along with hyperparameter tuning with final outputs for classification being 68% and number of clusters being k=6.Future scope of this paper is to try using Random Forests or Random Trees to improve accuracy of the model |
| 22 | Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project | 2017 |  | Manal Alghamdi et al. PLoS One. 2017. | The paper talks about the different machine learning approaches in medical research, it takes up the dataset ofThe Henry Ford exercise testing (FIT) project. The dataset contained 62 attributes classified into four categories: demographic characteristics, disease history, medication use history, and stress test vital signs. They developed an Ensembling-based predictive model using 13 attributes that were selected based on their clinical importance, Multiple Linear Regression, and Information Gain Ranking methods. The results showed that logistic regression classifier achieves the highest performance (69.1% for G1 the first attribute; 68.9 for G2) while the J48 has the lowest performance.4 |
| 23 | Diabetes detection using deep learning algorithms | 2020 |  | Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India | In this paper, input HRV signals employing deep learning architectures of CNN, LSTM and its combinations. They achieve a high accuracy of 95.7% employing CNN 5-LSTM architecture with SVM using 5-fold cross-validation. The Electrocardiograms (ECG) of 20 people each from the diabetes and normal group were collected for 10 min with people lying down in a relaxed supine position. The heart rate [time series data](https://www.sciencedirect.com/topics/computer-science/time-series-data) is derived from ECG signals using Pan and Tompkins algorithm. |
| 24 | Predictive modelling and analytics for diabetes using a machine learning approach. | 2019 |  | Department of computer science and engineering, school of engineering sciences, Jamia Hamdard, New Delhi, India. | In this paper the model is implemented using supervised machine learning techniques in R for Pima Indian diabetes dataset for understanding patterns. They have studied 5 different models based upon linear kernel support vector machine(SVM-linear), radial basis kernel support vector machine (SVM-RBF), k-nearest neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR) algorithms to detect diabetes in female patients. Dataset of female patients with minimum twenty one year age of Pima Indian population has been taken from UCI machine learning repository. Support vector machine (SVM) is used in both classification and regression. In SVM model, the data points are represented on the space and are categorized into groups and the points with similar properties falls in same group. The experimental results suggested that all the models achieved good results; SVM-linear model provides best accuracy of 0.89 and precision of 0.88 for prediction of diabetes as compared to other models used. On the other hand *k-*NN model provided best recall and F1 score of 0.90 and 0.88. |

| 25 Diabetic Retinopathy using morphological operations and machine learning | 2017 | 2015 IEEE International Advance Computing Conference (IACC) |  | This paper focuses on detection of Diabetes Retinopathy which leads to blindness in 80-85% diabetic patients. For automated detection, novel two-step hierarchical binary classification is used. For classification purposes GMM, SVM, KNN and ADABOOST methods are used. They take 30 top features like area, variance of Ired channel, Igreen channel, I sat of object, major and minor axis length, Mean pixels for Igreen, Ired and Intensity, solidity etc. After detecting exudates and micro-aneurysms in color fundus image, the features get extracted. Splat, GLCM (Gray Level Co-occurrence Matrix) and calculated are applied to SVM and KNN classifier. SVM classifier gives better results than KNN classifier. |
| --- | --- | --- | --- | --- |
| 26 A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques | 2020 | 2019 1st International Informatics and Software Engineering Conference (UBMYK) |  | This paper takes the PIMA indian diabetes dataset which consists of 768 instances with 8 features. The data was divided into training(60%) and test(40%). The performance of the proposed method was evaluated in terms of overall accuracy (OA), Kappa Coefficient (KC), precision (P), recall (R), and f-measure (F). The SVM was efficient for data classification and the random forest algorithm was used for various different classification. The convolution neural network was also used. The SVM classification produced and over all accuracy 73.94 %, precision: 62.56%, recall: 45.82%, F-measure: 51.93% and KC. The RF produced more effective classification, The overall accuracy obtained for this classier was 79.26%, precision: 84.36%, recall: 62.74%, f-measure: 70.93% and kappa: 0.556. |
| 27 Prediction and diagnosis of diabetes mellitus — A machine learning approach | 2015 | 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) |  | In this paper different decision support systems are introduced using various data mining algorithms for assisting medical experts. The local dataset of different places in kerala is collected. Machine learning calculations extricate valuable information and concealed examples from the dataset for forecast and determination. Support vector machine was used to diagnose PIMA indian diabetes dataset, but it showed that the adaptive neuro fuzzy Inference showed better accuracy for the diagnosis of diabetes. The system they used focuses on Adaboost algorithm which is implemented using 4 stages. The result they got in the end was The computer information system with AdaBoost-decision stump classifier provides an accuracy of 80.729 % for predicting diabetes with a very low value of error rate |
| 28 Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Technique | 2020 | IEEE Access (volume 8) |  | The authors talk about a novel technique based on NDDM to accommodate irregular and sparsely sampled EMRs data with the objective to overcome the analytical problems. The dataset acquired is from CPCSSN. It contains unique data of about 170,000 patient,  Spanning 13 years timeframe. The final dataset comprised 775 (61.03%) female and 1143 (38.96%) male and among them 584 (23.49%) are diabetic patients. The experimental results showed that the INDDM performed better than the NDDM. As a secondary analysis, a logistic regression analysis is also conducted to assess the significant p-value of each risk factor with the incident of T2DM.In this study, a series of experiments are designed to analyze and compare INDDM with standard NDDM for handling irregular and sparsely sampled EMRs data as input for prognostic prediction task. All experiments are conducted using the previously described data obtained from CPCSSN. |
| 29 Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection | 2018 | 2018 International Conference on Innovation in Engineering and Technology (ICIET) |  | In this paper the PIMA indian dataset is taken and the authors are using the neural network model in this process. They have three different steps in this section such as data recovery, feature selection, and M.L.P. Classifier. As a first step, data recovery techniques are applied by replacing the missing data with the mean value for making the dataset complete for building a model. Then, they do the feature selection process which is done to find the features that have the most impact on the risk factor identification. Lastly, a suitable number of hyper-parameters are selected that works well for this data-set. They apply the K-means algorithm in this section after selecting the features. We also use different machine learning classifiers to compare the outcome. The k-means algorithm effectively reduces noise from the data. They replace a lot of data missing in the pima dataset by taking the mean value or the most likely value of the given feature. They select certain features only relevant to the study and discard the ones that don't have an impact on the prediction model. The accuracy level on the train dataset was 86.73% and 85.15%. For future scope the authors want to try it on actual patients from hospitals |
| 30 Predictive analysis of diabetic patient data using machine learning and Hadoop | 2017 | 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) |  | In this paper various different types of data mining techniques are integrated to form a predictive analysis. In this work we are using the machine learning algorithms in the Hadoop-Map Reduce environment for the prediction of the types of diabetes that are widespread, related complications and accordingly treatment can be provided. On the basis of this analysis, the system will be able to provide a competent solution to the early diagnosis of patients' risk level. The predictive analysis system architecture comprised various phases like data collection, missing values imputation, pattern discovery, pattern matching and result analysis. The Pima Indians Diabetes Data set is used for the experimentation. In order to discover patterns from the data set, they have given Pima diabetes data set to C4.5 algorithm. There are four attributes selected to reduce complexity of results. Attributes are selected on the basis of following criteria. The algorithm generates total 57 rules from standard data set. According to the highest information gain ratio, attribute plasma glucose (0) has been selected as a root node. In future work pattern matching will be employed by applying these discovered patterns on testing data set to predict diabetic prevalent and risk levels associated with it. |