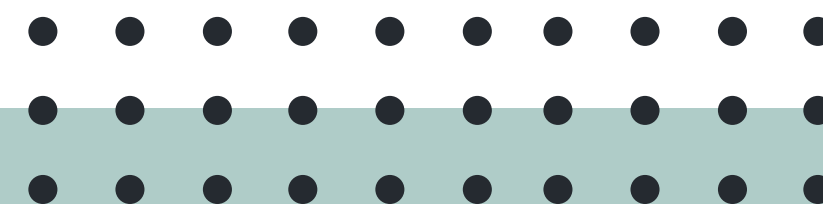


DIABETES

A Deep Learning
Approach to
Diabetes Diagnosis






What is Diabetes?



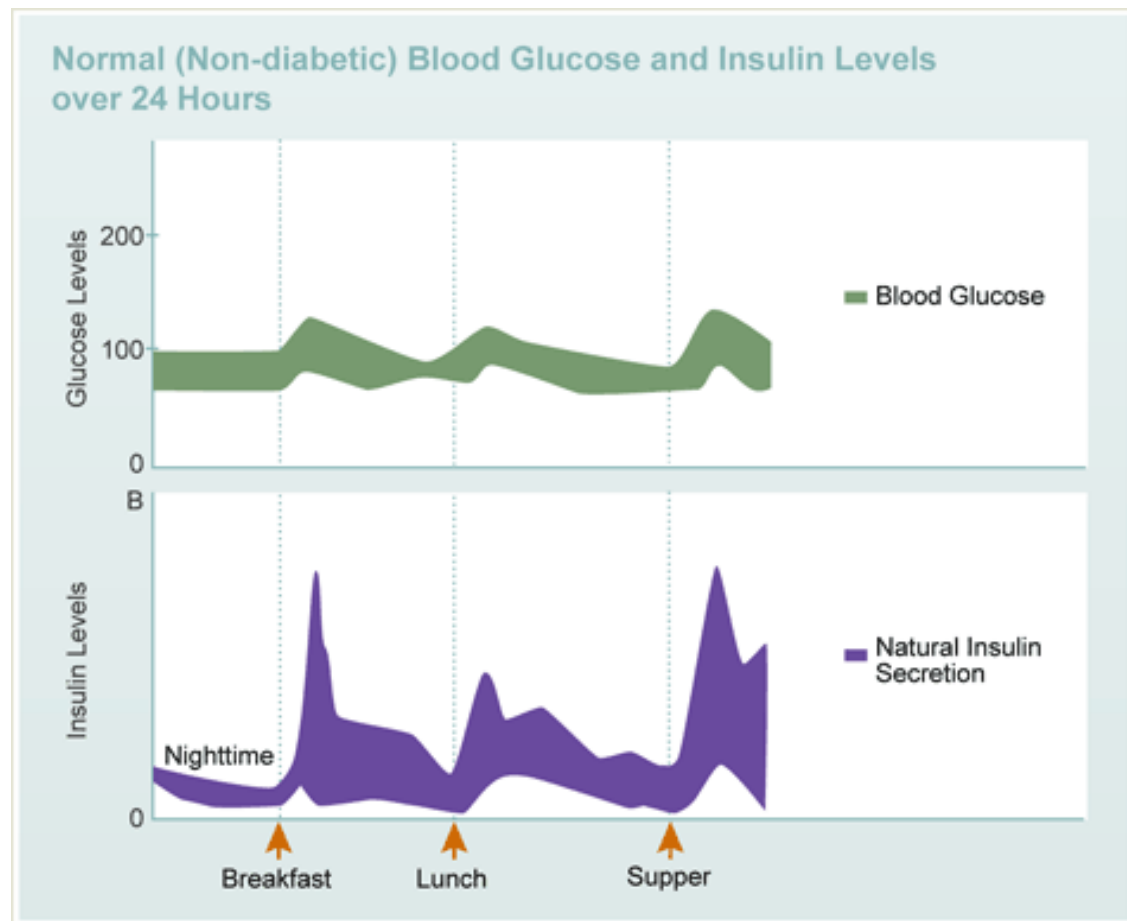
Diabetes mellitus is the medical and scientific term given for an unusual and unhealthy High blood sugar level while it is often tested and found through various methods, one such reference we will use to describe it will be through a glucose tolerance test.

- The Average/Normal level – 140 mg/dL (7.8 mmol/L)
 - An Elevated level (Indicating prediabetes) – 140-199 mg/dL (7.8 mmol/L – 11.0 mmol/L)
 - Diabetic Level (Duration of 2 hours) – 200 mg/dL (11.1 mmol/L)
- 



Problem

While readings may show that diabetes exist, it is the situation and circumstance of the cause for the diabetic reading that is most important. This is due to the diabetes being separated into 2 types.



Type 1 Diabetes

A lifelong condition where the body's immune system attacks and destroys the cells that produce insulin

01

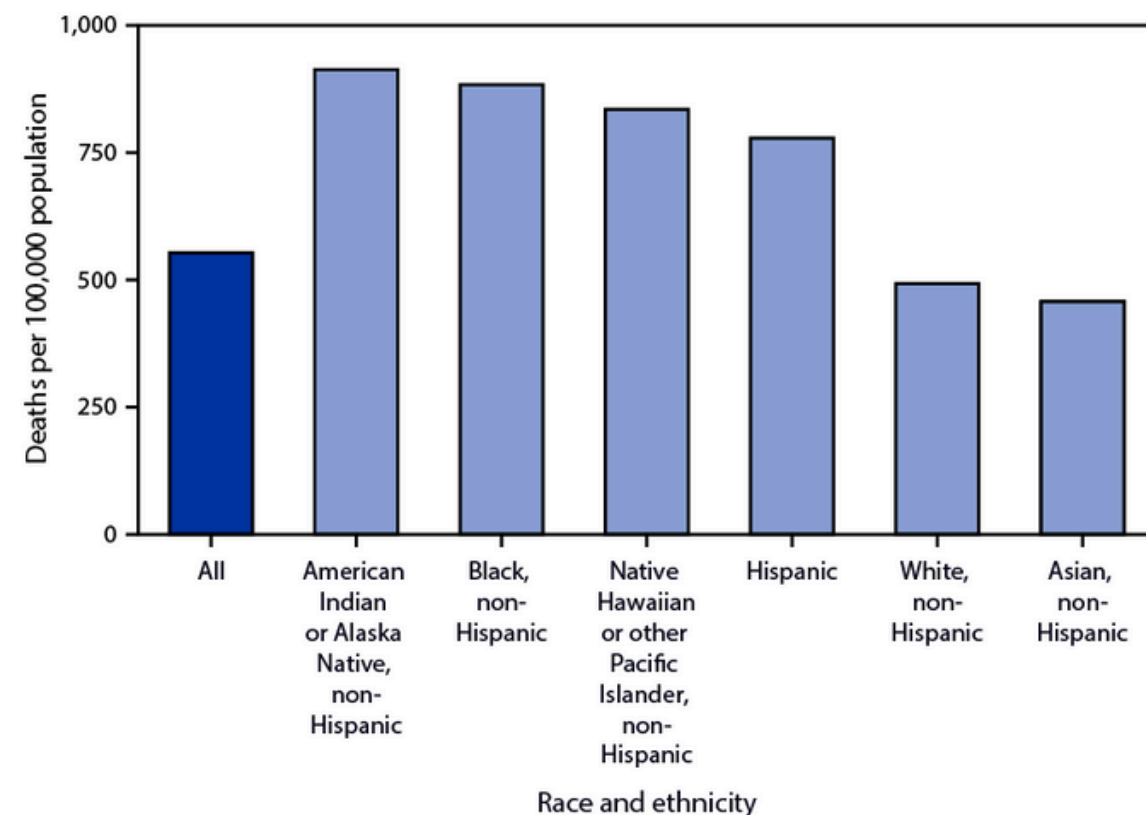
Type 2 Diabetes

Wherein the body is unable to produce the required amount of insulin.

02

Diabetes and its Fatality Rate

- Estimations range that 4.2 million deaths across the ages of 20 to 79 are attributed towards diabetes
- This is furthered by the estimate of diabetes contributing to 11.3% of deaths globally.
- Ranging from 6.8% (Lowest) in Africa to 16.2% (Highest) in the Middle East and North Africa



As per 2020, the following is data visualized and collected from the United States.

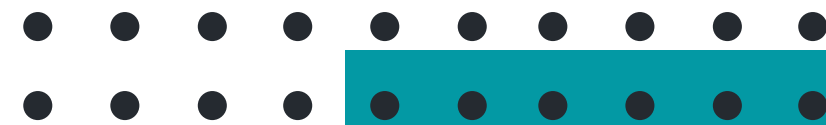
(Saeedi, P., Salpea, P., Karuranga, S., Petersohn, I., Malanda, B., Gregg, E. W., Unwin, N., Wild, S. H., & Williams, R. (2020). Mortality attributable to diabetes in 20-79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, 162, 108086.)



Diagnosis of Diabetes

Diabetes is primarily detected through blood tests that measure blood sugar levels. Some common methods include:

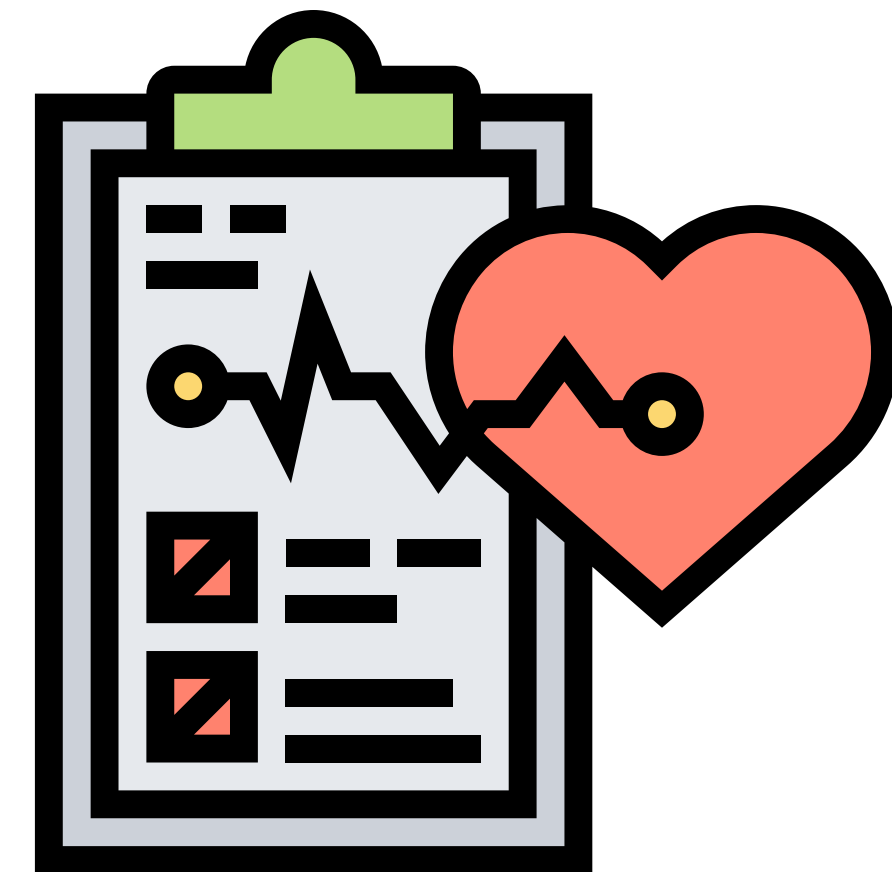
- Fasting Plasma Glucose (FPG)
- A1C Test (HbA1c)
- Oral Glucose Tolerance Test (OGTT)
- Random Plasma Glucose Test



Early Diagnosis Importance

Historically, type1 diabetes is commonly known and is expected to be a disease that primarily affects those in the ages of 10-19, thus, consequently, diagnosis and efforts have been largely focused towards younger populations. however, it has also been prevelant among and far above this age range.

- Preventative Measures
- Minimize Complications
- Improved Quality of Life
- Reduced Healthcare Costs






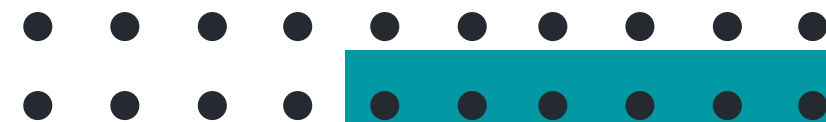
Misdiagnosis

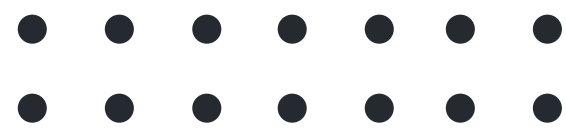


Where there are multiple methods of detecting diabetes, such as through invasive blood glucose laboratory tests and glucometers which are standard for glucose monitoring at hospitals and homes.



There is also disadvantages and underlying costs to use them, such as a strict requirement for the skillset and equipment, to the prohibitive costs, the time consumed and the comfortability of the process via the pain associated with it. These causes together lead to the opportunity for misdiagnosis.

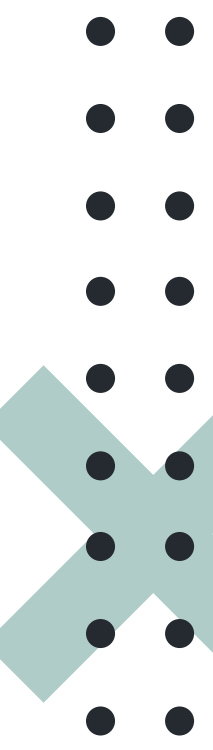




Common use Dataset and Parameters

The Pima Indians Diabetes Dataset is a well-known dataset used in machine learning for classification tasks, particularly related to predicting diabetes.

- The Pima Indians Diabetes dataset is a collection of data on 768 Pima Indian women aged 21 or older.
- Compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).
- The dataset allows researchers to develop and test machine learning models to predict whether a patient is likely to have diabetes based on the provided diagnostic measurements.






Methodology



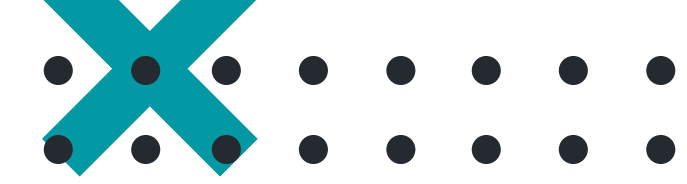
Implemented basic ML models as referenced from the base research paper:



Zhang, Z., Ahmed, K. A., Hasan, M. R., Gedeon, T., & Hossain, M. Z. (n.d.). A deep learning approach to diabetes diagnosis. arXiv preprint arXiv:2403.07483.

The ML Models are implemented using Scikit-Learn and TensorFlow





About the Base Research Paper

Summary:

Problem Addressed:

- Traditional diabetes diagnostic methods are invasive and costly.
- Existing machine learning models (CkNN, GRNN) perform poorly with imbalanced data.

Proposed Solution:

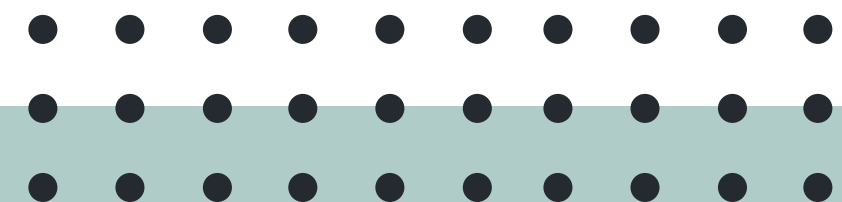
- A non-invasive diabetes diagnosis method using a Back Propagation Neural Network (BPNN).
- Incorporates batch normalization, data resampling, and normalization for class balancing.

Key Benefits:

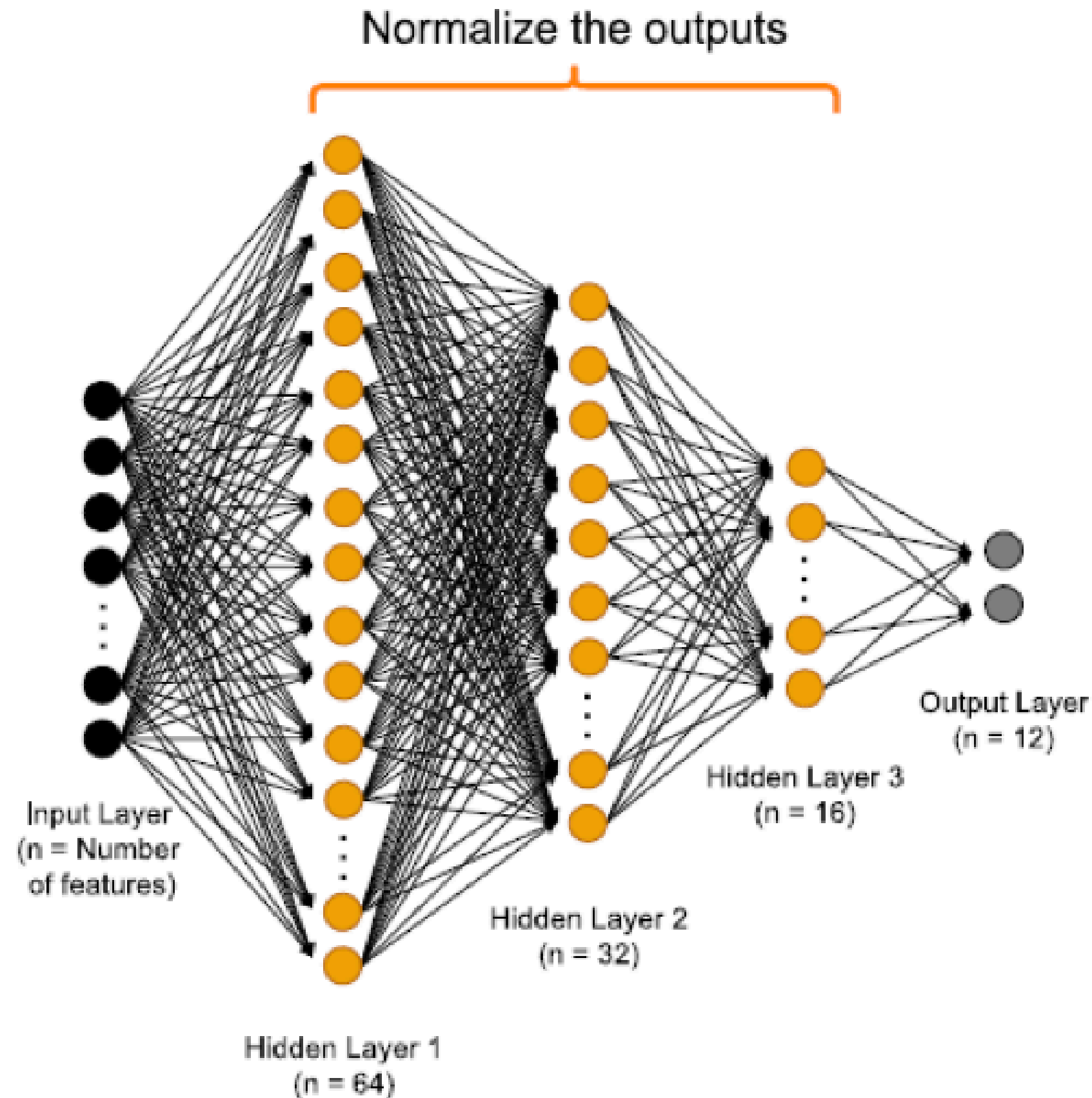
- Overcomes performance limitations of traditional machine learning methods.
- Significant improvements in overall accuracy, sensitivity, and specificity.

Experimental Results:

- Pima diabetes dataset: Achieved 89.81% accuracy.
- CDC BRFSS2015 dataset: Achieved 75.49% accuracy.
- Mesra Diabetes dataset: Achieved 95.28% accuracy.



Back Propagation Neural Network (BPNN)





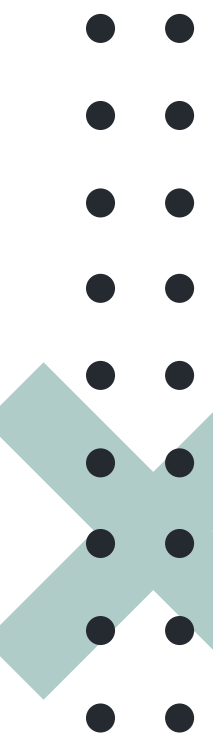
Integration with the dataset

To evaluate the the effectiveness of the BPNN model on the dataset, we compare it with some of the other statistical learning methods, deep learning methods and existing methods done by similar and related works.

Dataset used:

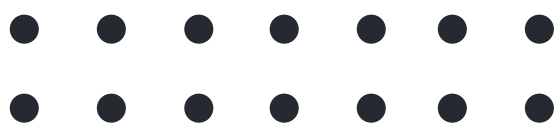
PIMA Indians Diabetes Dataset:

(<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>)
by the US National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [3,19,20] and available at the University of California Irvine Machine Learning Repository (Zhang et al.)





Data Preprocessing



Quick Look at our Dataset

	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





Imbalanced Data




When a dataset has a lot more examples of one outcome (like healthy) compared to another (like diabetic), it's called imbalanced.

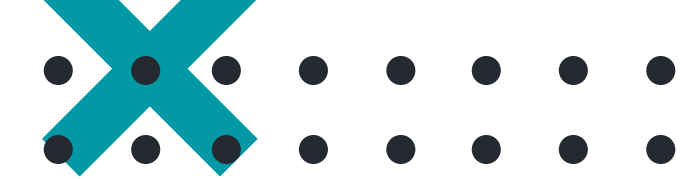
Issue:

- Imbalanced datasets lead to biased models that favor the majority class.
- This means the model might be great at predicting healthy people but bad at identifying diabetes.

Our dataset was skewed in favour of non diabetic outcomes. So we under sample our data

(Bunkus, O., BunkutÃ©, L., & Sruoga, V. (2021). An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. Applied Sciences, 12(8), 3928. <https://doi.org/10.3390/app12083928>.)



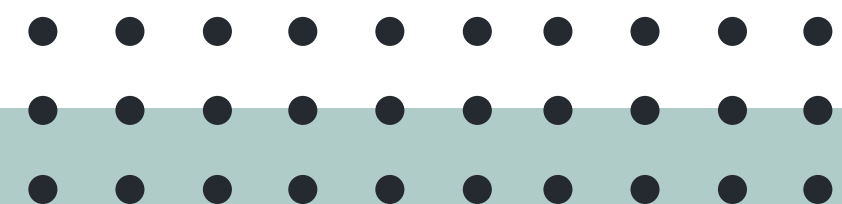


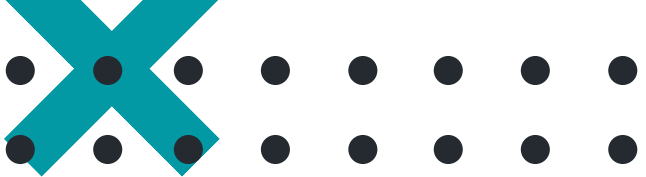
Data balancing

- Making the dataset more even.
- Reduces bias and improves the model's performance for both outcomes.

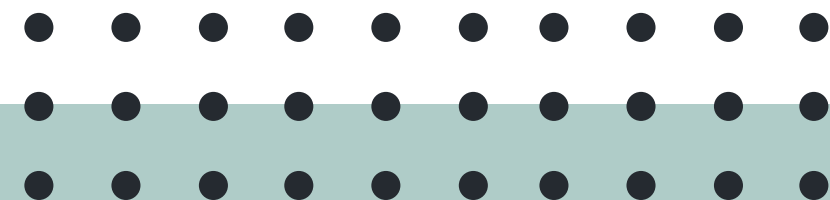
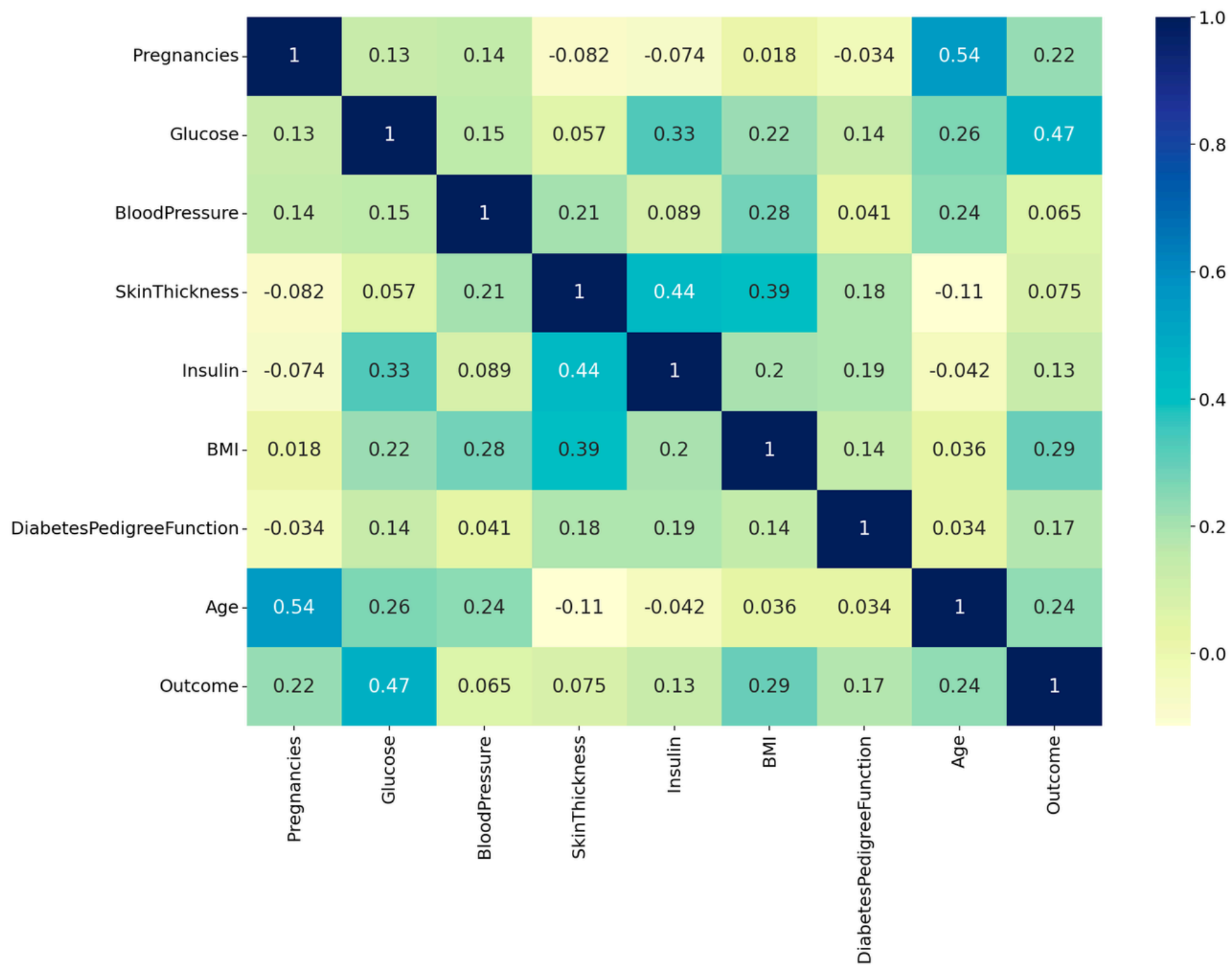
The Pima Indians Diabetes dataset:

- **Problem:** The original dataset had significantly more people without diabetes (500) than with diabetes (268). This imbalance can affect machine learning models.
- **Solution:** Undersampling was applied to balance.
- **Drawback:** Undersampling discards some data from the original dataset.





Covariance Matrix






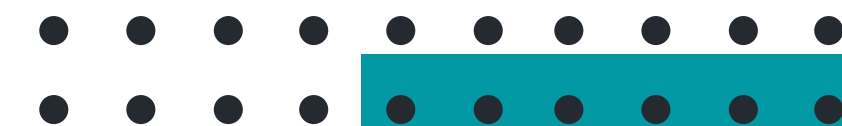
Data Scaling

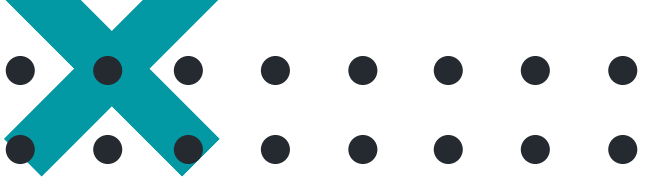


- Many machine learning algorithms, especially for classification, use distances between data points to make predictions.
- Uneven scales lead to problems
- By scaling:
 - Improved performance
 - Faster training
- Data Scaling Techniques:
 - **Normalization:** Scales features to a specific range (e.g., 0 to 1).
 - **Standardization:** Scales features to have a mean of 0 and a standard deviation of 1.

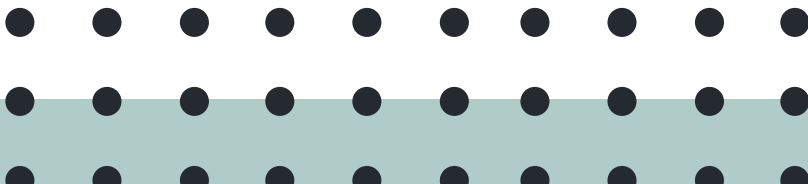
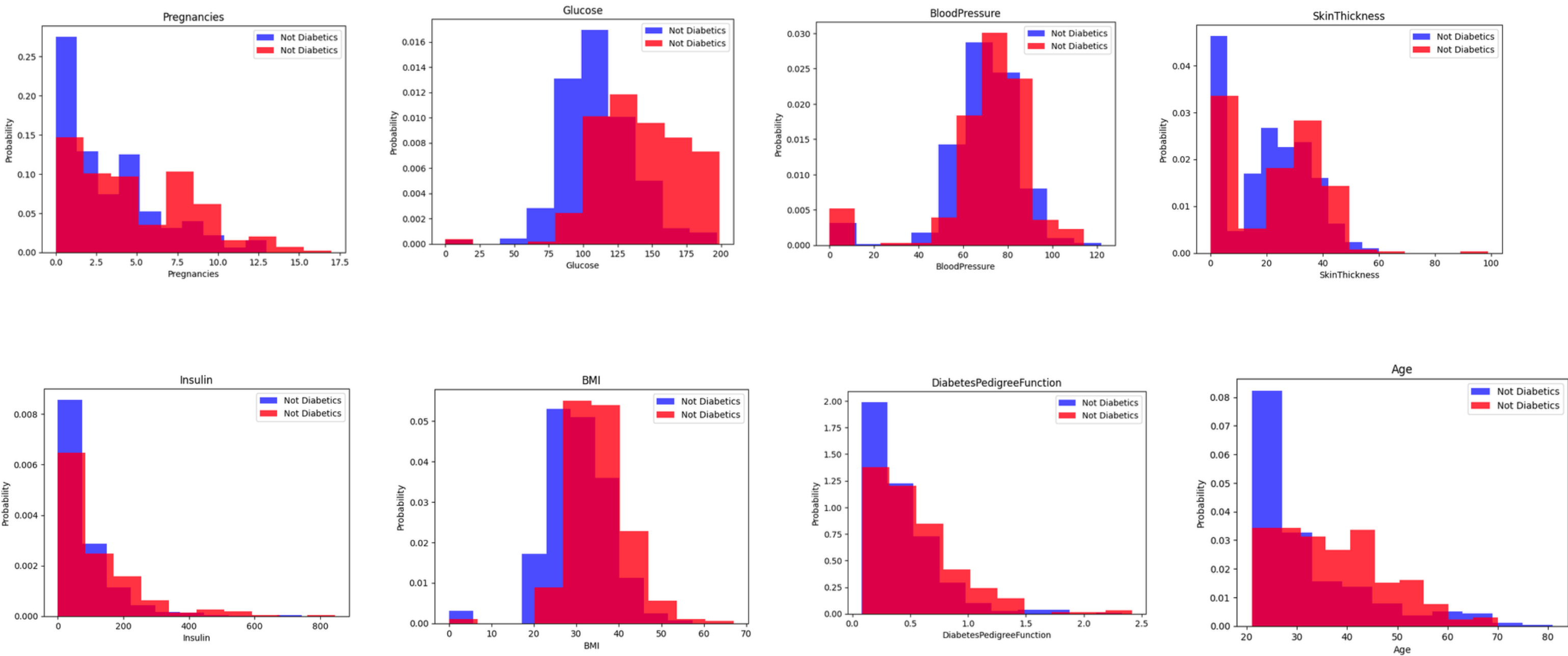


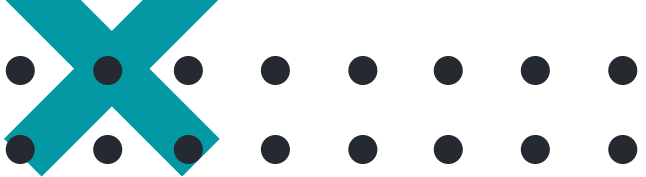
(Ahsan, MM, Mahmud, MAP, Saha, PK, Gupta, KD, & Siddique, Z. (2021).
Effect of Data Scaling Methods on Machine Learning Algorithms and Model
Performance. Technologies, 9(3), 52.
<https://doi.org/10.3390/technologies9030052>.)



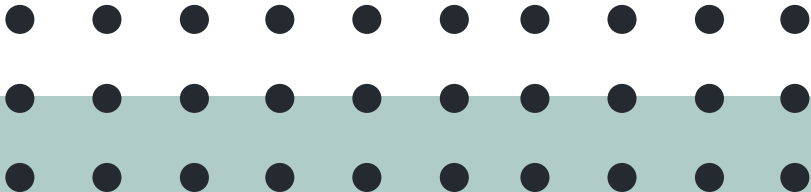
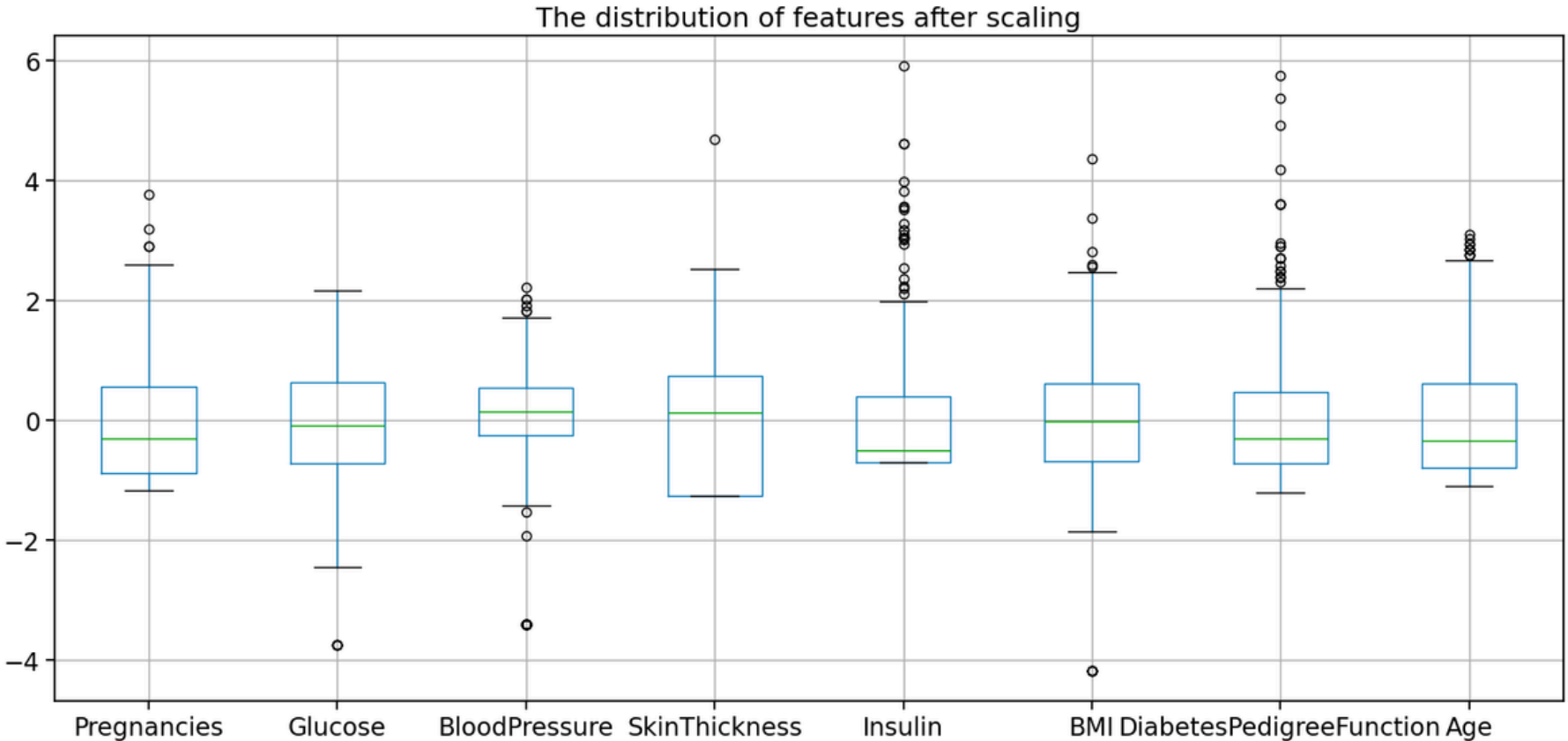
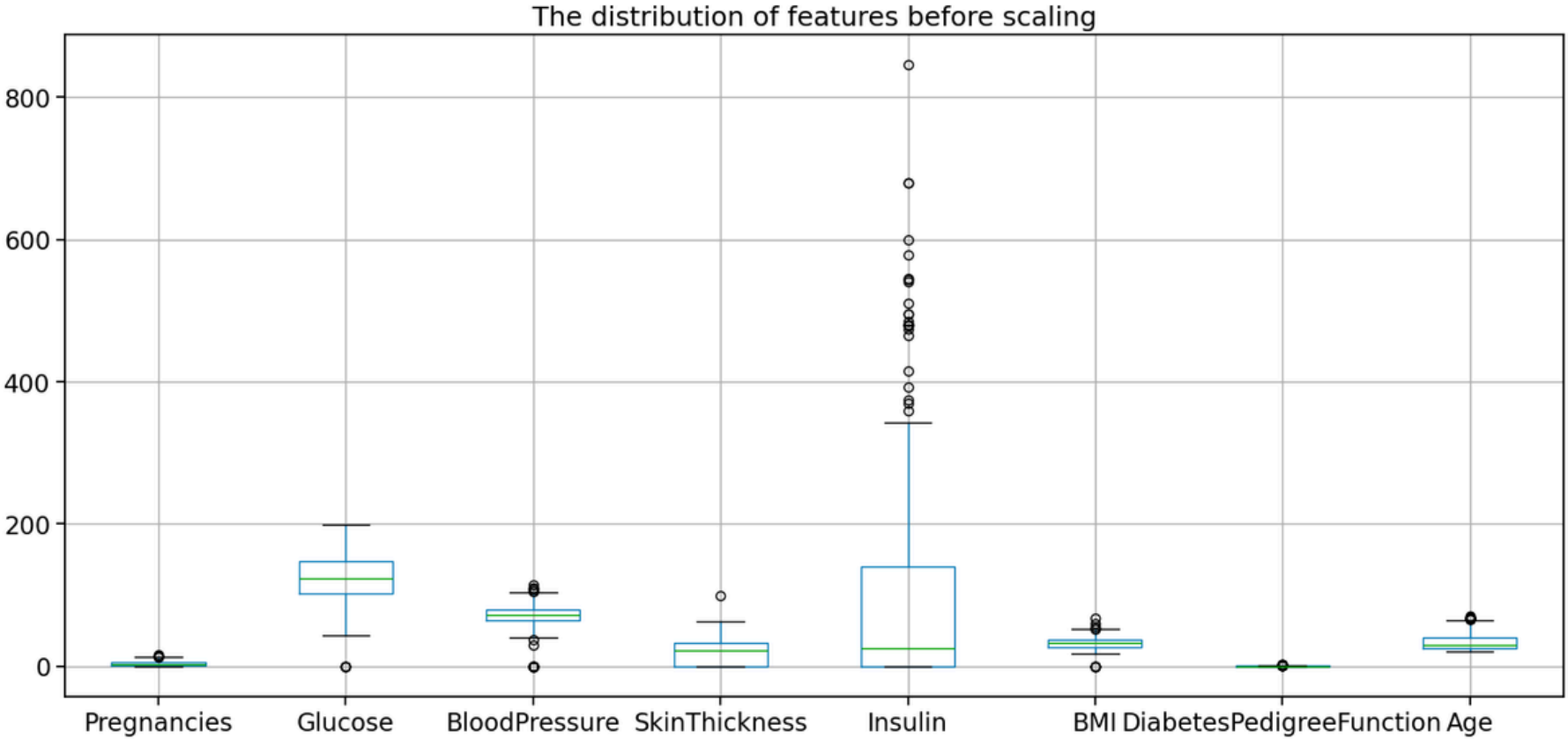


Data Visualisation





Box Plot





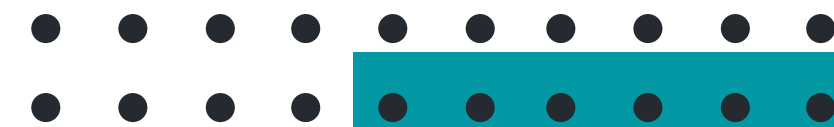
Hyperparameter Tuning

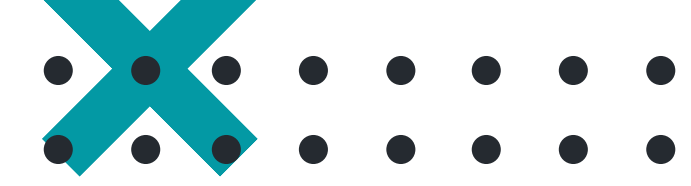


Machine learning models have parameters that are learned from data. Hyperparameters are settings that control the learning process itself.

Grid Search is a method for finding the best combination of hyperparameters for a model.

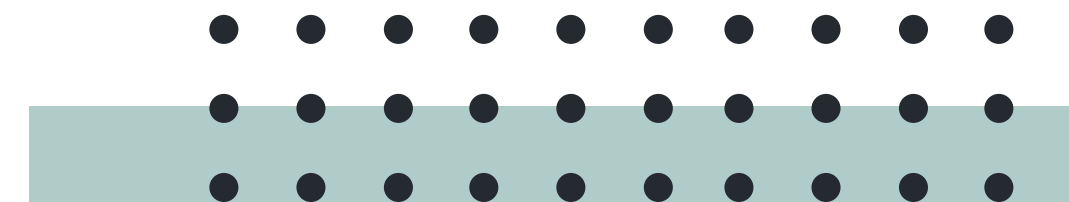
- Defining a grid of possible values for each hyperparameter.
- Training the model with every combination of hyperparameters in the grid.
- Evaluating the model's performance on a validation set for each combination.
- Choosing the combination of hyperparameters that leads to the best performance.





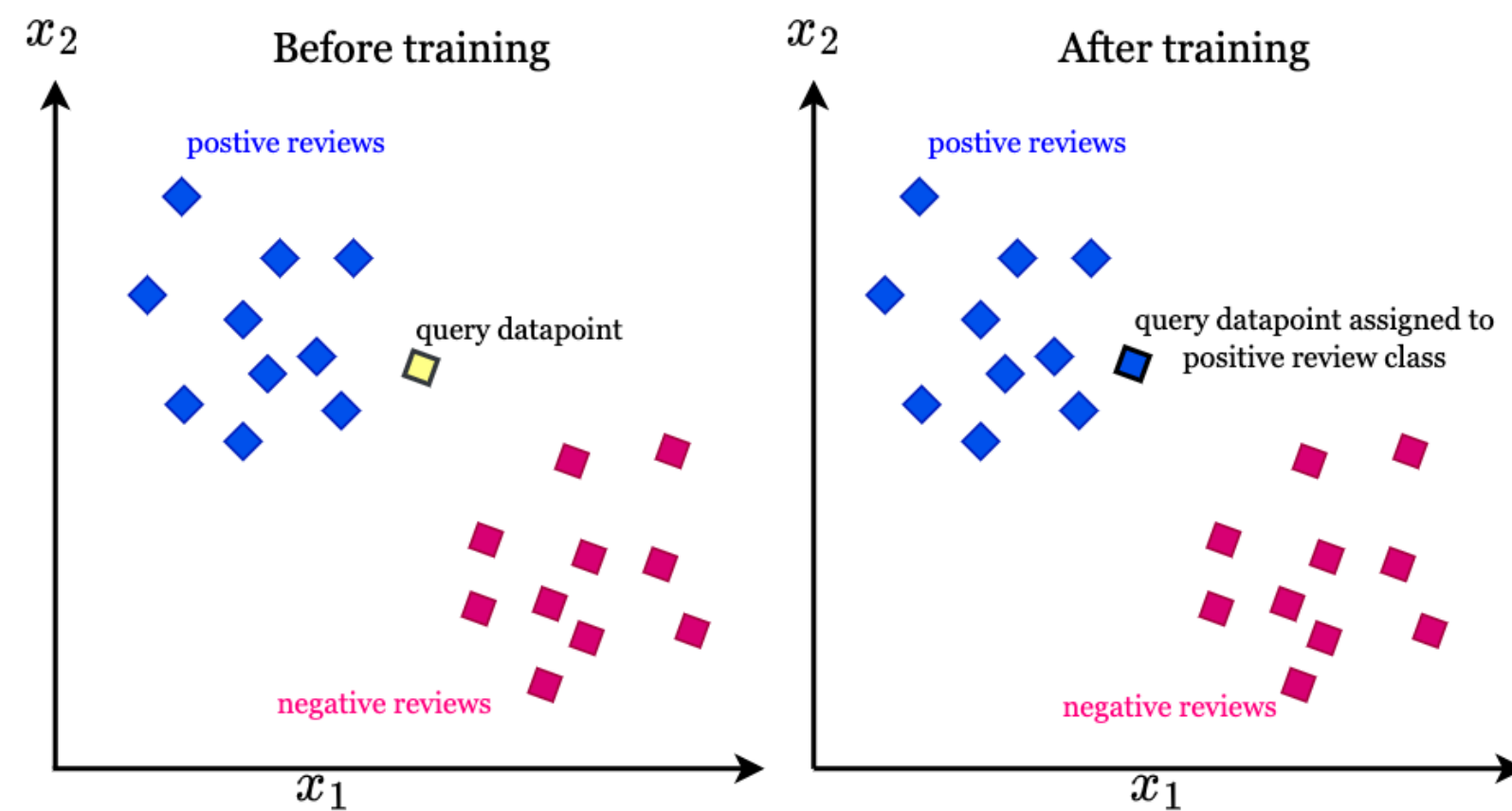
Validation & ML Models Used

We used a train-test split of 80% to 20%



KNN

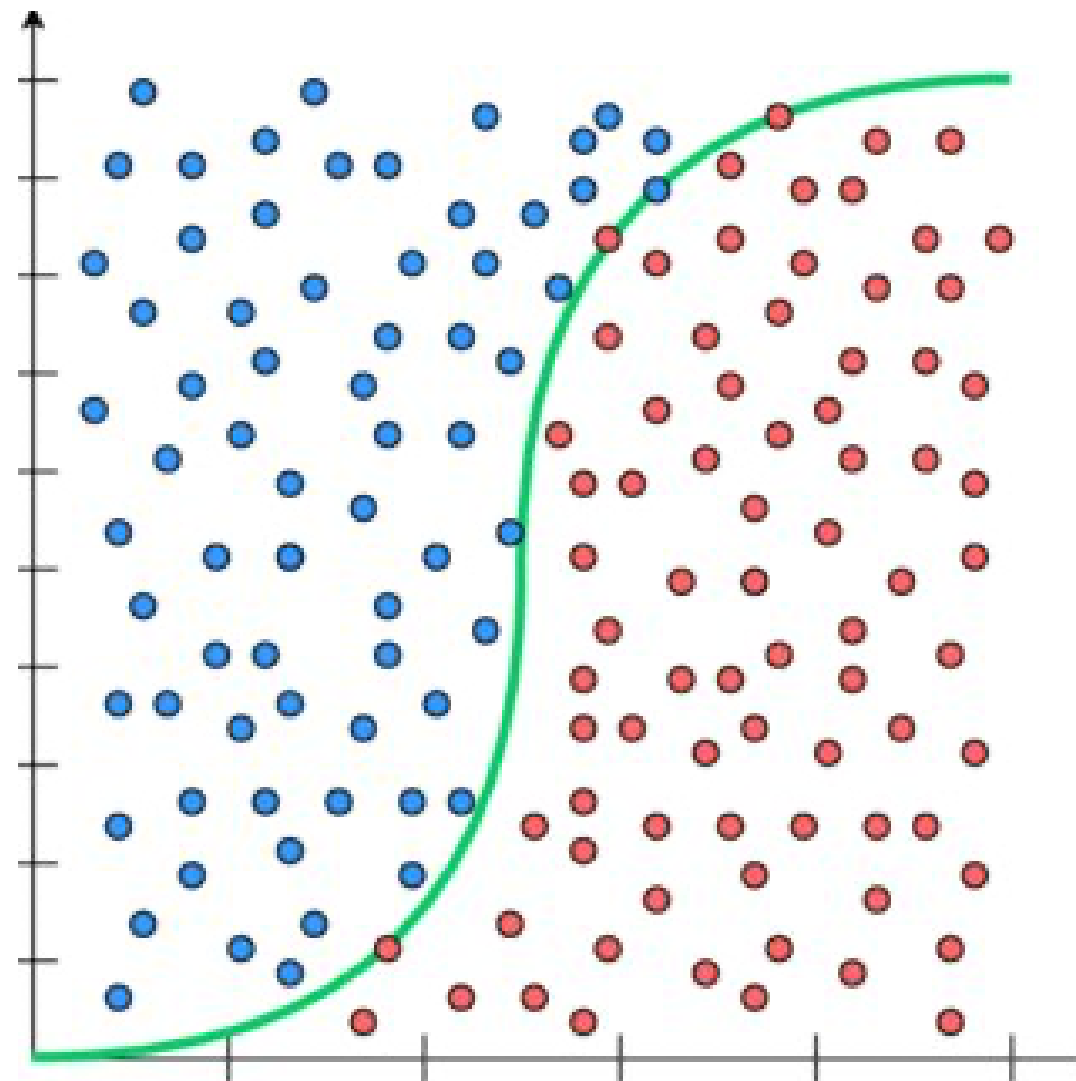
- A powerful classification algorithm used in pattern recognition
- Stores all available cases and classifies new cases based on similarity based on a similar measure





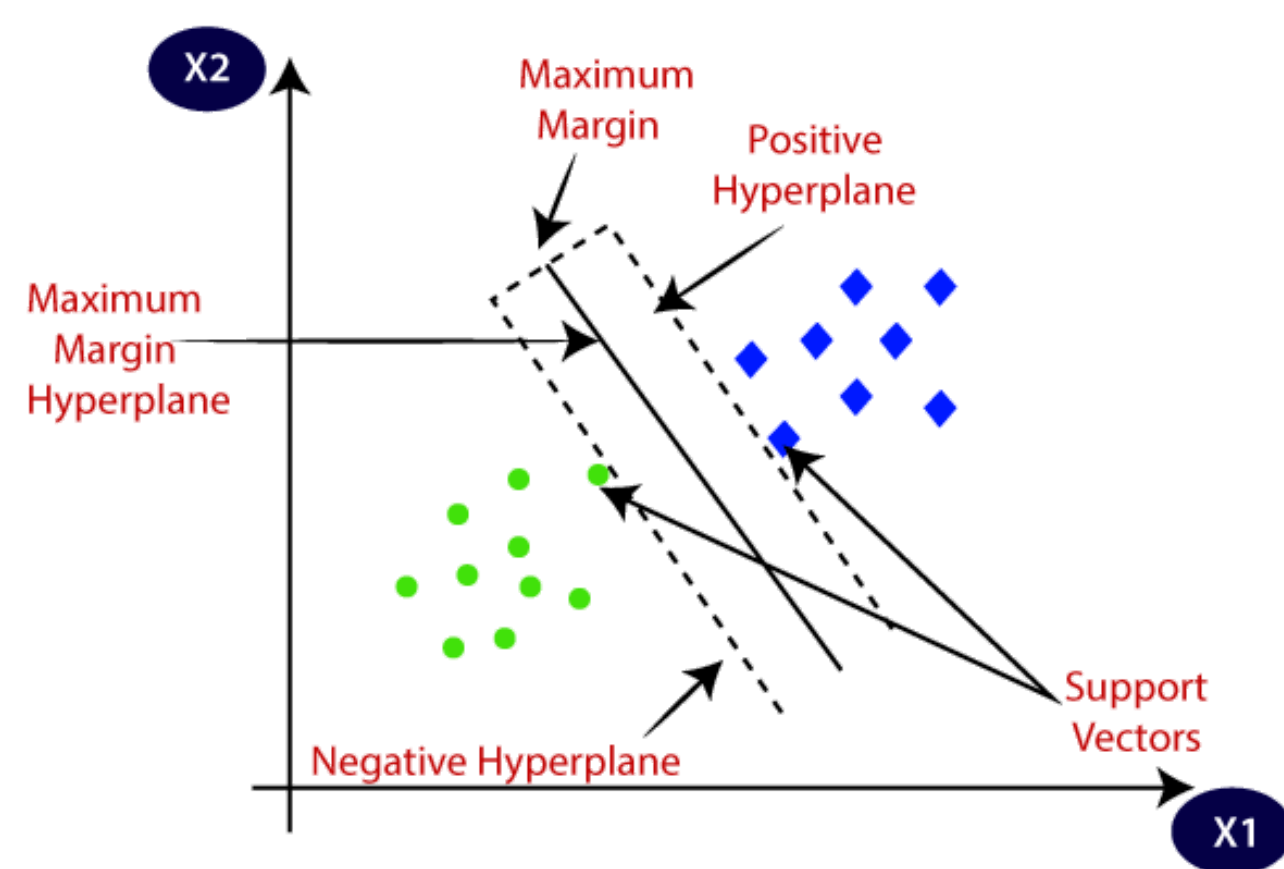
Logistic Regression

- Classification algorithm used for predicting the probability of categorical dependent variable
- The dependent variable is mainly binary variable i.e {0,1}



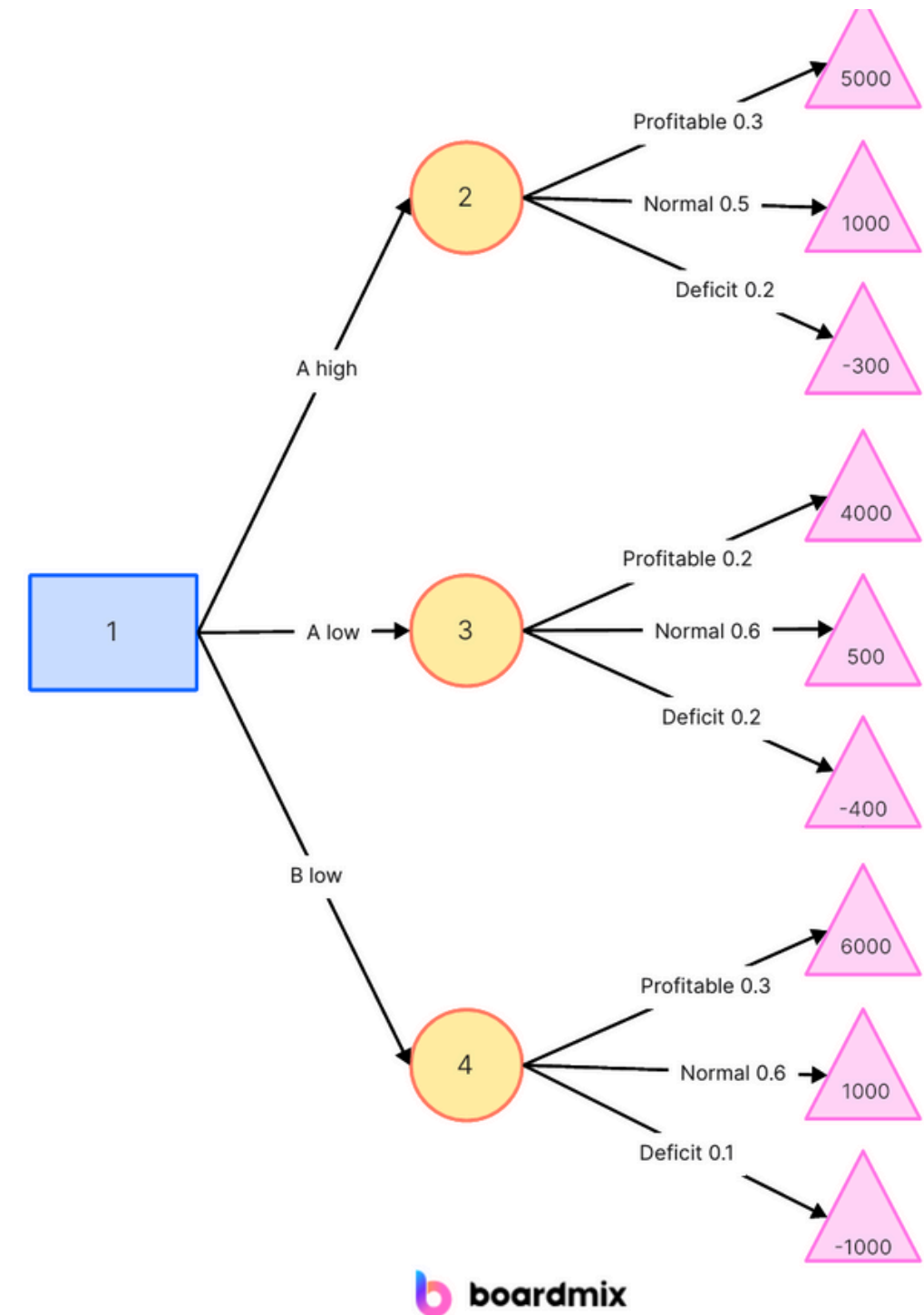
Support Vector Machine

- Make decisions by recursively splitting the input space into regions.
- the splitting is done with the use of an hyperplane and values on the margine, are the support vector



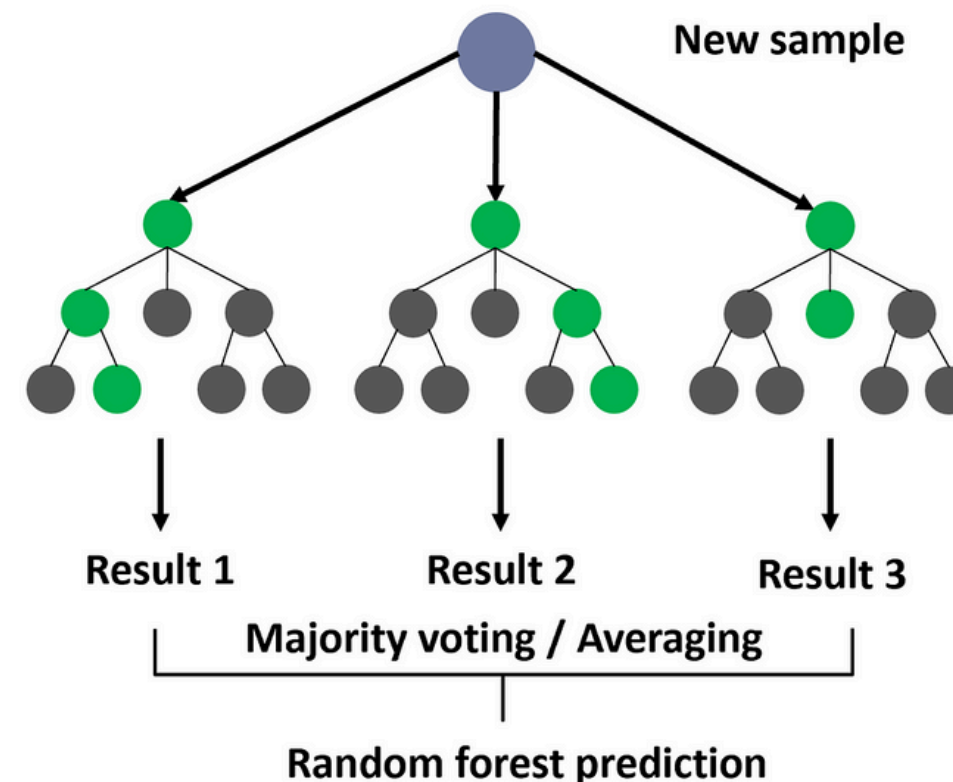
Decision Trees

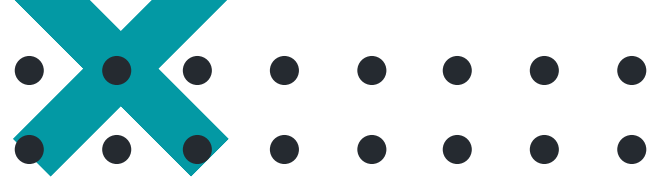
- supervised learning algorithm used for both classification and regression tasks.
- It works by recursively splitting the data into subsets based on the value of input features, creating a tree-like model of decisions and their possible consequences.



Random Forest Classifier

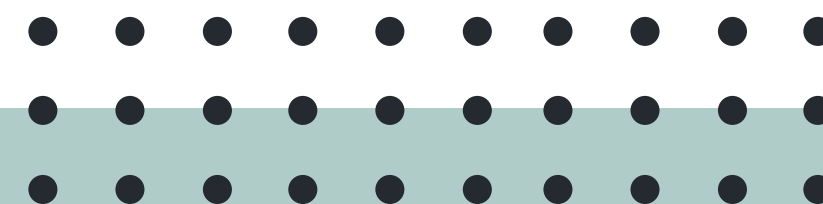
- ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- It is a powerful and versatile machine learning algorithm capable of performing both regression and classification tasks.





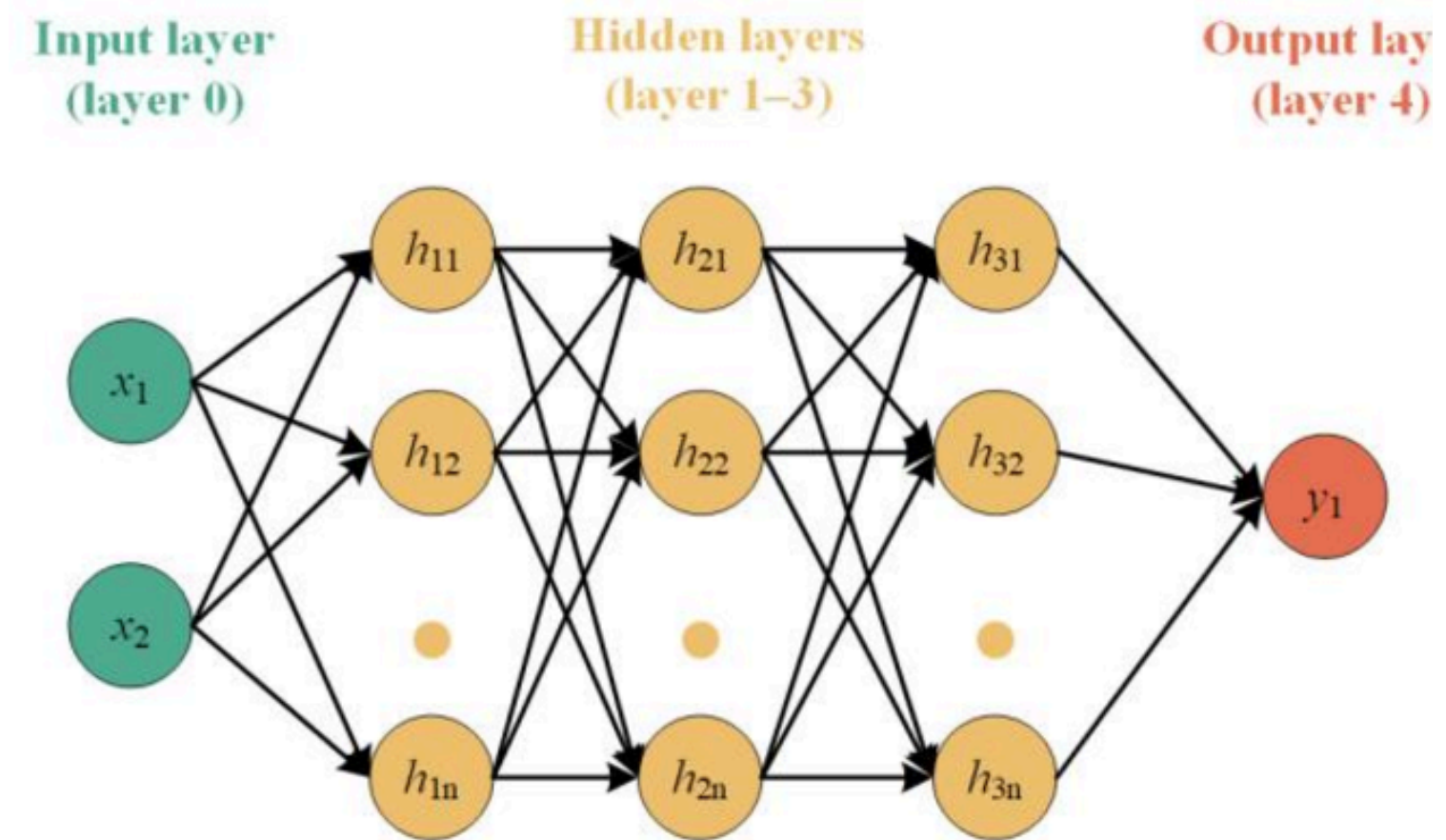
Boosting

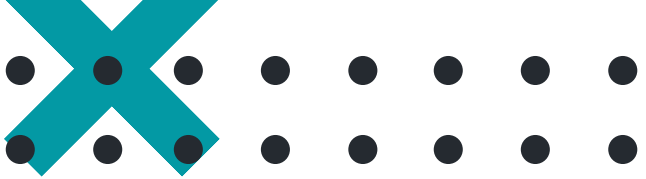
- ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a powerful and versatile machine learning algorithm capable of performing both regression and classification tasks.
- **Adaboost:** Sequentially train weak learners, updating sample weights to focus on misclassified examples. Combine the weak learners' predictions using weighted voting, where weights are based on each learner's accuracy.
- **XGBoost:** Iteratively add trees to correct errors of the previous ensemble using gradient descent. Minimize a regularized objective function combining loss function and model complexity.



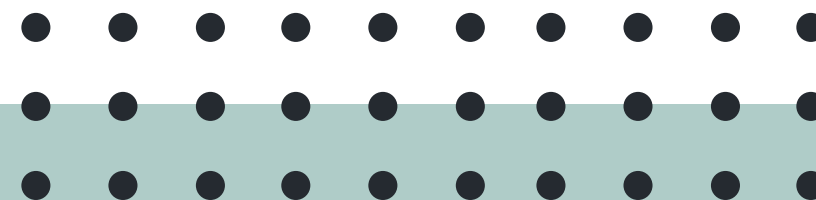
BPNN

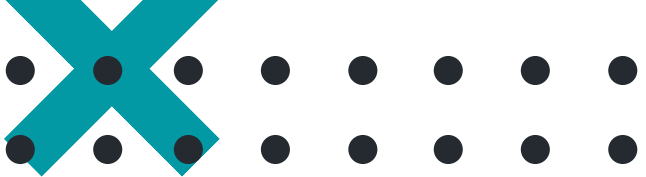
- Backpropagation Neural Network (BPNN) is a type of artificial neural network where information flows forward through layers during the feedforward phase and backward during the backpropagation phase to adjust weights based on errors.
- It's implemented by training data through multiple iterations, minimizing errors using optimization algorithms like stochastic gradient descent (SGD), and updating weights accordingly.



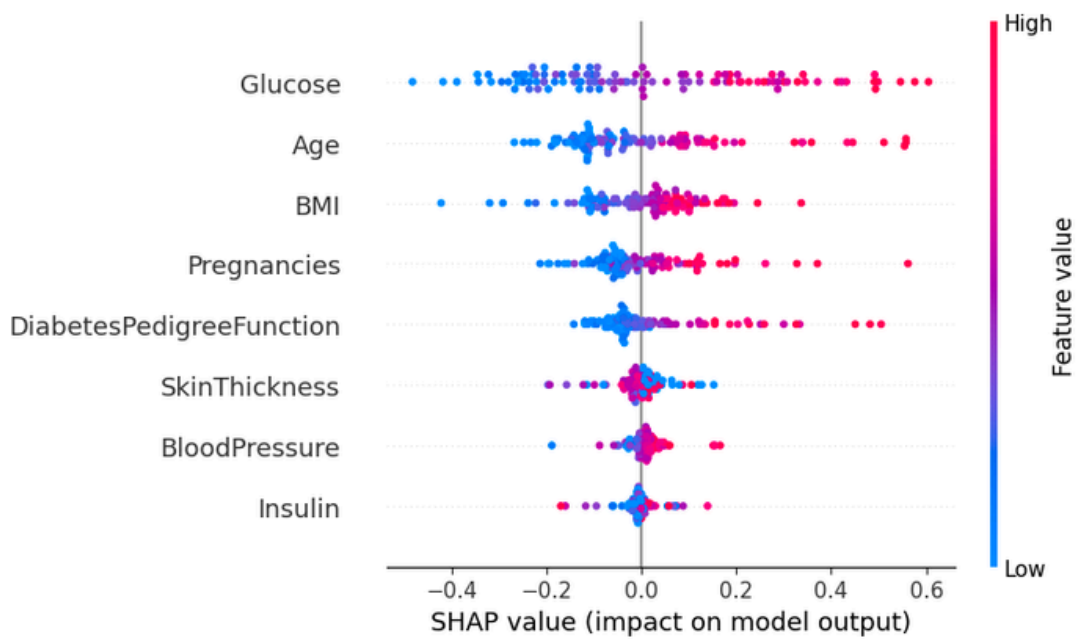


Results

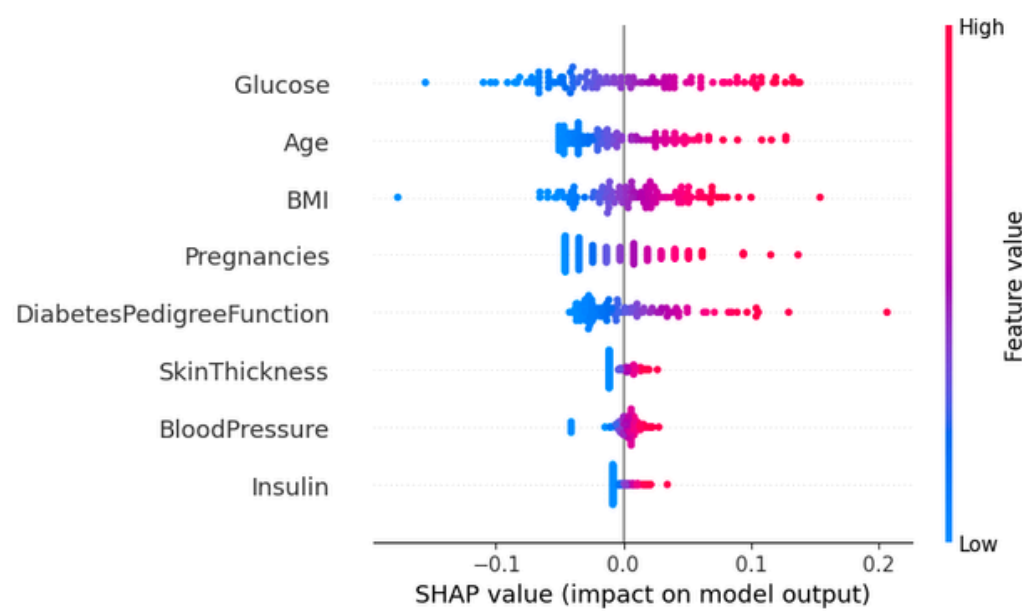




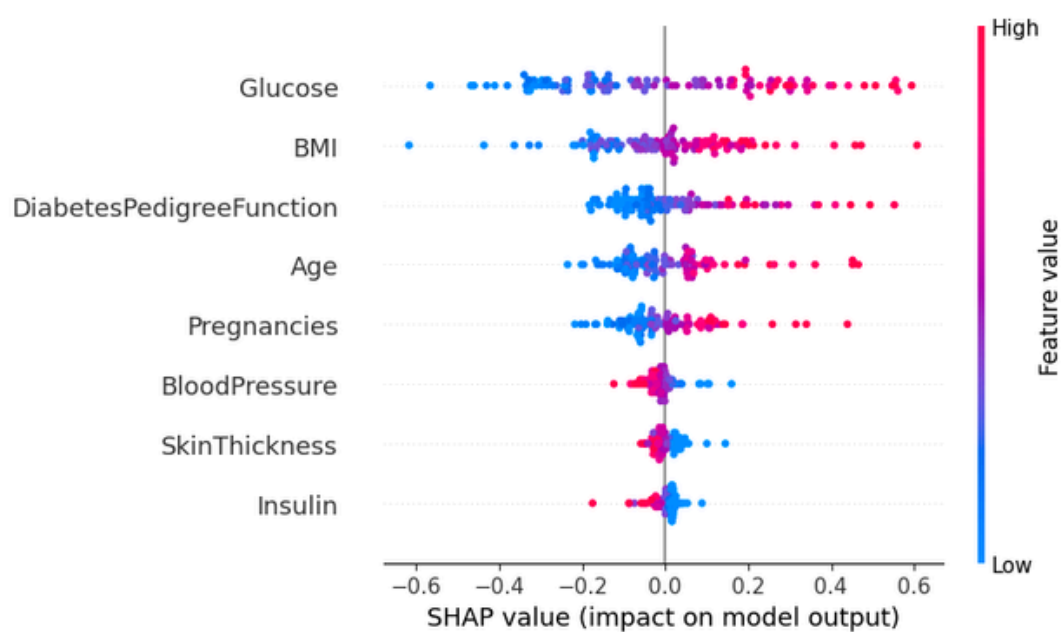
SHAP Values



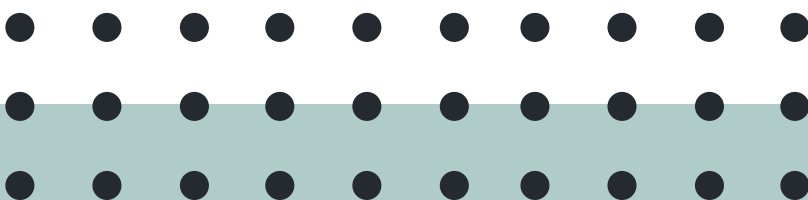
KNN



Logistic Regression



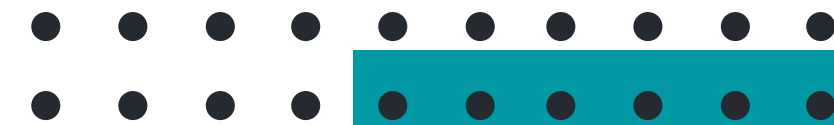
Support Vector
Machine

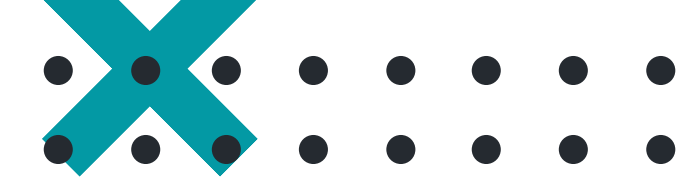




Models and Accuracies

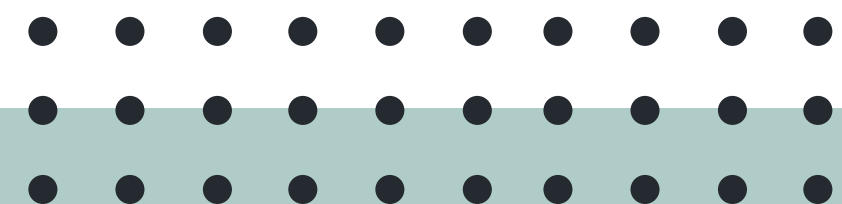
- K-Fold Neural Network - 74.53%
- Logistic Regression - 76%
- Support Vector Machine (SVM) - 79%
- Decision Trees - 79%
- Random Forest Classifier - 74%
- XG Boost - 87%
- ADA Boost - 80%
- BPNN - 82%





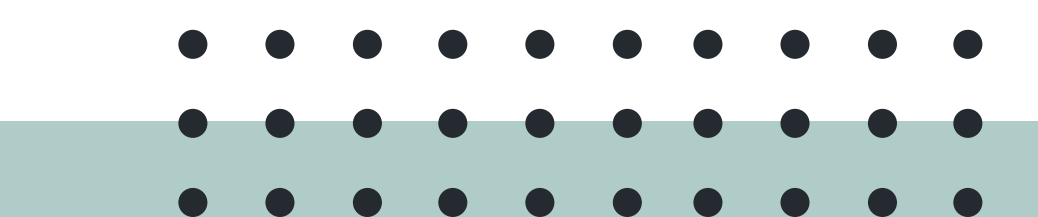
References

- Bunkus, O., BunkutÃ©, L., & Sruoga, V. (2021). An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. Applied Sciences, 12(8), 3928. <https://doi.org/10.3390/app12083928>.
- Ahsan, MM, Mahmud, MAP, Saha, PK, Gupta, KD, & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), 52. <https://doi.org/10.3390/technologies9030052>.
- Sharma, V. (2022). A study on data scaling methods for machine learning. International Journal for Global Academic & Scientific Research, 1(1), 31-42.
- Pareek, J., & Jacob, J. (2021). Data compression and visualization using PCA and T-SNE. In Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019 (pp. 327-337). Springer Singapore.






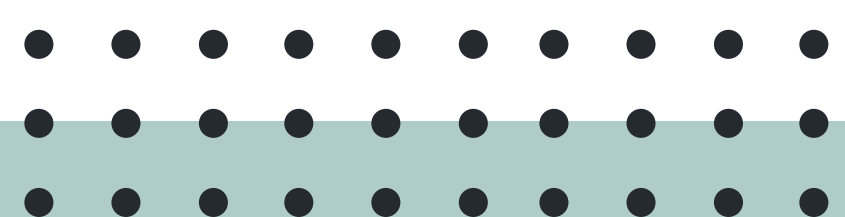
References

- Alibrahim, H., & Ludwig, S. A. (2021, June). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In 2021 IEEE Congress on Evolutionary Computation (CEC) (pp. 1551-1559). IEEE.
 - Liu, Y., & Liao, S. (2014). Preventing over-fitting of cross-validation with kernel stability. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14 (pp. 290-305). Springer Berlin Heidelberg.
 - Zeyu Zhang, Khandaker Asif Ahmed, Md Rakibul Hasan, Tom Gedeon, Md Zakir Hossain. (2024, March). A Deep Learning Approach to Diabetes Diagnosis.
 - Saeedi, P., Salpea, P., Karuranga, S., Petersohn, I., Malanda, B., Gregg, E. W., Unwin, N., Wild, S. H., & Williams, R. (2020). Mortality attributable to diabetes in 20-79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, 162, 108086.
- 



References



- Shivkumar, S., & Magliano, D. J. (2023). Type 1 diabetes misdiagnosis as type 2 diabetes in adults: Time for action. *The Lancet Regional Health - Europe*, 29, 100385. <https://doi.org/10.1016/j.lanepe.2023.100385>
 - Fauci, A. S., Lane, H. C., & Redfield, R. R. (2023). HIV/AIDS. In W. C. Watson (Ed.), *Harrison's Principles of Internal Medicine* (20th ed.). McGraw-Hill Education.
<https://www.ncbi.nlm.nih.gov/books/NBK555976/>
 - Polak, R., & Eliakim, R. (2019). Nonalcoholic fatty liver disease in non-obese individuals: A call for personalized treatment. *Digestive Diseases and Sciences*, 64(4), 939-943. <https://doi.org/10.1007/s10620-019-05538-7>
 - Christobel, Y.A., Sivaprakasam, P. A new classwise k nearest neighbor (cknn) method for the classification of diabetes dataset. *International Journal of Engineering and Advanced Technology* 2(3), 396–400 (2013)
 - Bennett, P. H., Burch, T. A., & Miller, M. (1971). Diabetes mellitus in American (Pima) Indians. *The Lancet*, 298(7716), 125-128.
[https://doi.org/10.1016/S0140-6736\(71\)92303-8](https://doi.org/10.1016/S0140-6736(71)92303-8)
- 
- 



Thank You!



- Aaron Maricar - Senior Secondary
- Chaitanya Patil - B.Tech. (2nd Year)
- Rudraksh Srivastava - B.Tech. (2nd Year)
- Paavan Kumar S - B.Tech. (3rd Year)
- Monish N S - B.Tech. (2nd Year)



Our Code: <https://github.com/ArtConnoisseur/NTU-Team-Research>

