Problem Statement & Model Selection:-

Problem Statement

- Predict whether a patient has diabetes based on medical attributes
- Binary classification problem

Training Model Used

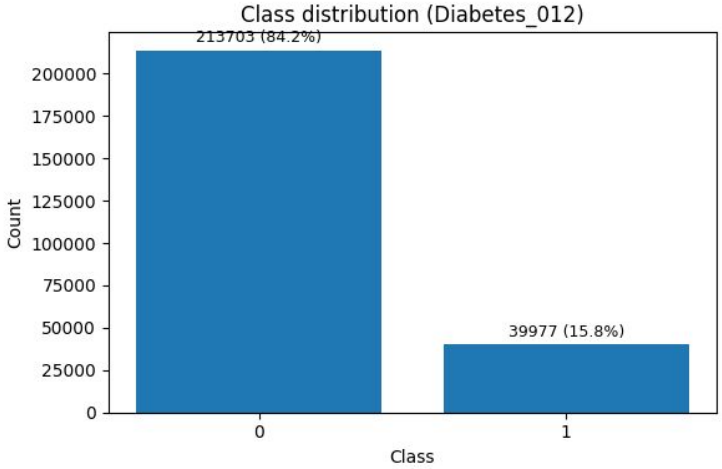
- **Logistic Regression**
- Supervised learning algorithm for binary classification

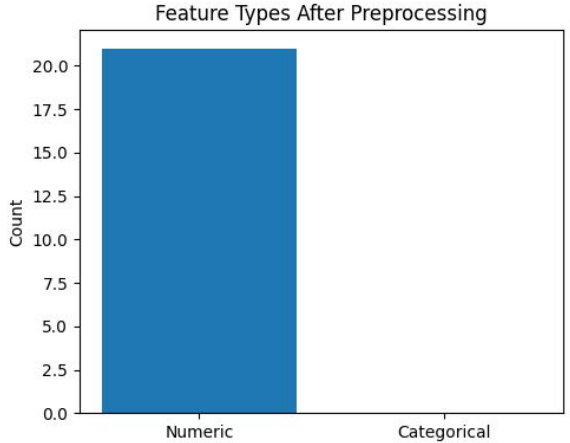
Why Logistic Regression

- Interpretable baseline model
- Efficient for medium-to-large datasets
- Works well with standardized numerical features
- Suitable for demonstrating effects of randomized scaling techniques

Why Supervised Learning and Binary Classification

- Dataset contains labeled examples (X,y)
- Ground truth outcomes are available for every patient
- Enables direct optimization of prediction accuracy using known labels
- Clinical outcome is categorical with two possible states
- Logistic Regression naturally models probability of a binary event
- Output is interpretable as diabetes risk ($P(y=1 \mid X)$)





Asymptotic Running Time Analysis:-

Model Training Complexity

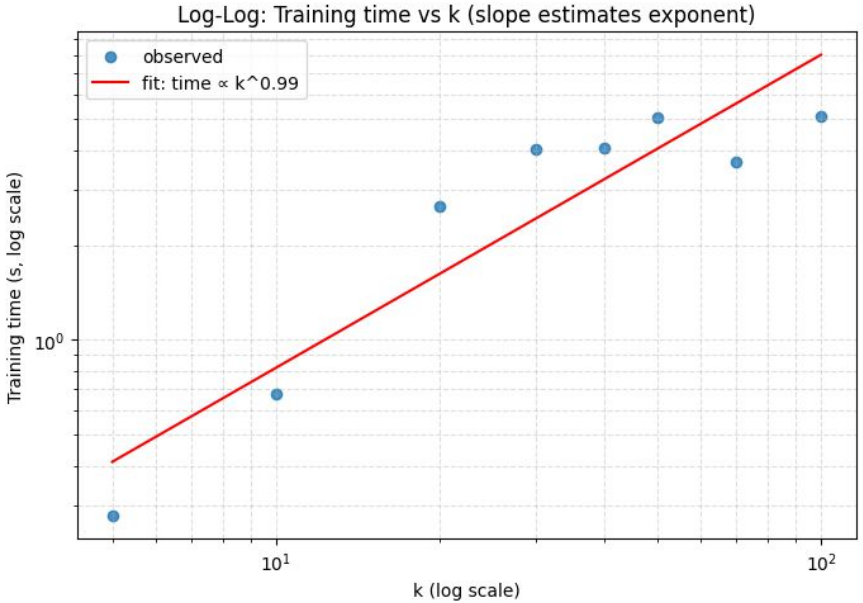
- Let:
  - n = number of samples
  - d = number of features

Logistic Regression

- Trained using iterative optimization (e.g., gradient descent)
- Per-iteration time complexity: O(nd)

With Randomized Scaling

- After Random Projection to k dimensions, where $k \ll d$
- Time complexity becomes: O(nk)



Key Insight

- Reducing feature dimension significantly lowers computational cost per iteration
- Enables faster training on larger datasets

Model Accuracy Performance:-

Accuracy without Scaling

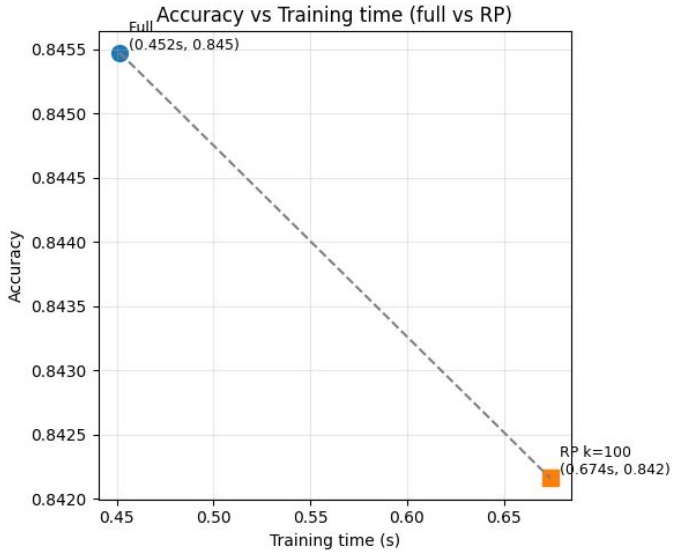
- Test Accuracy: **84.55%**

Accuracy with Random Projection

- Test Accuracy: **84.22%**

Observation

- Accuracy loss after scaling is minimal
- Random Projection preserves predictive performance while improving efficiency



Randomized Scaling Technique Used:-

Scaling Technique

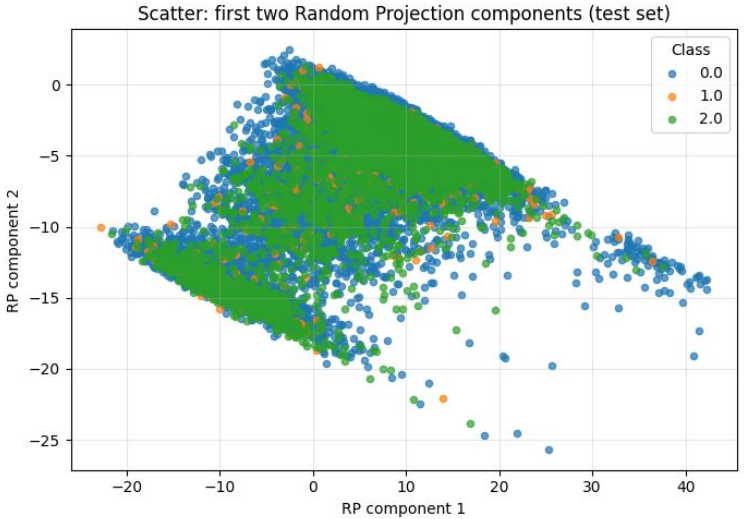
- **Random Projection**

What Random Projection Does

- Projects data from d dimensions to k dimensions using a random matrix
- Preserves pairwise distances in expectation (Johnson–Lindenstrauss property)
- Computationally efficient and data-independent

Why Random Projection

- Reduces feature dimension without heavy computation
- Suitable for large datasets and iterative models
- Does not require expensive covariance computations (unlike PCA)
- Simple to integrate with Logistic Regression



Model Training – Original vs Scaled Dataset:-

Original Dataset (No Scaling)

- Feature dimension: d = 21
- Model trained on full feature space
- Baseline training time and accuracy

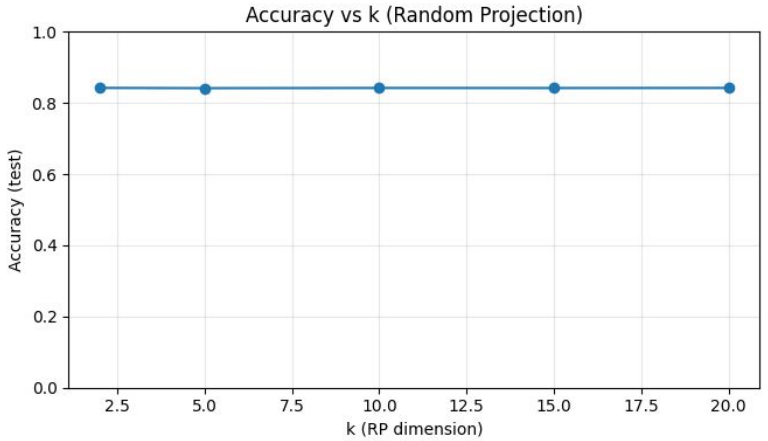
Scaled Dataset (Random Projection)

- Feature dimension reduced to k, where $k \ll d$
- Model trained on projected feature space
- Faster training with reduced dimensionality

Training Setup (Same for Both)

- Same train–test split
- Same Logistic Regression configuration
- Same evaluation metrics

	k	train_time	accuracy	roc_auc
0	2	0.604936	0.842380	0.592154
1	5	1.510291	0.841631	0.617624
2	10	2.964118	0.842163	0.735289
3	15	8.129744	0.841986	0.755646
4	20	14.122010	0.842242	0.767442



Accuracy Comparison:-

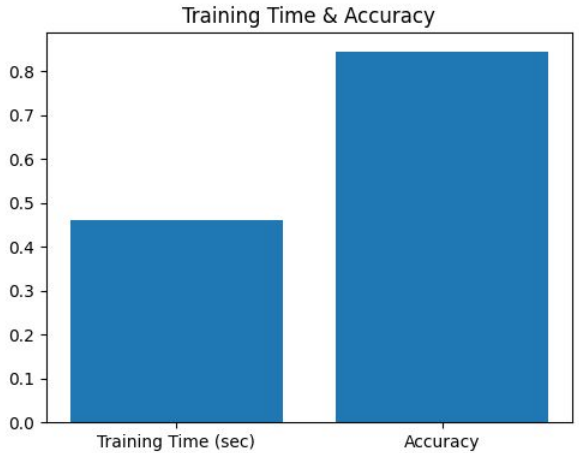
Model	Feature Dimension	Train Accuracy
Logistic Regression (Original)	(d)	**84.55%**
Logistic Regression + Random Projection	(k)	**84.22%**

Observations

- Test accuracy after scaling remains comparable
- Only minor accuracy variation after projection
- Randomized scaling maintains predictive performance

Key Takeaway

- Dimensionality reduction improves efficiency without significant loss in accuracy



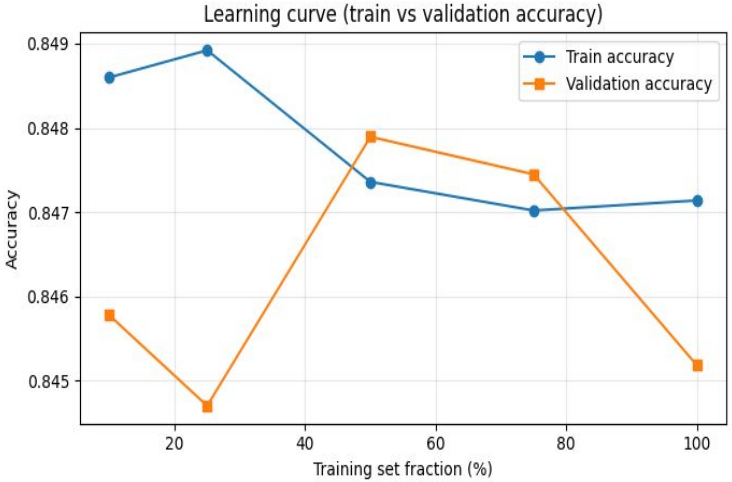
Training Time Improvement Using Randomized Scaling:-

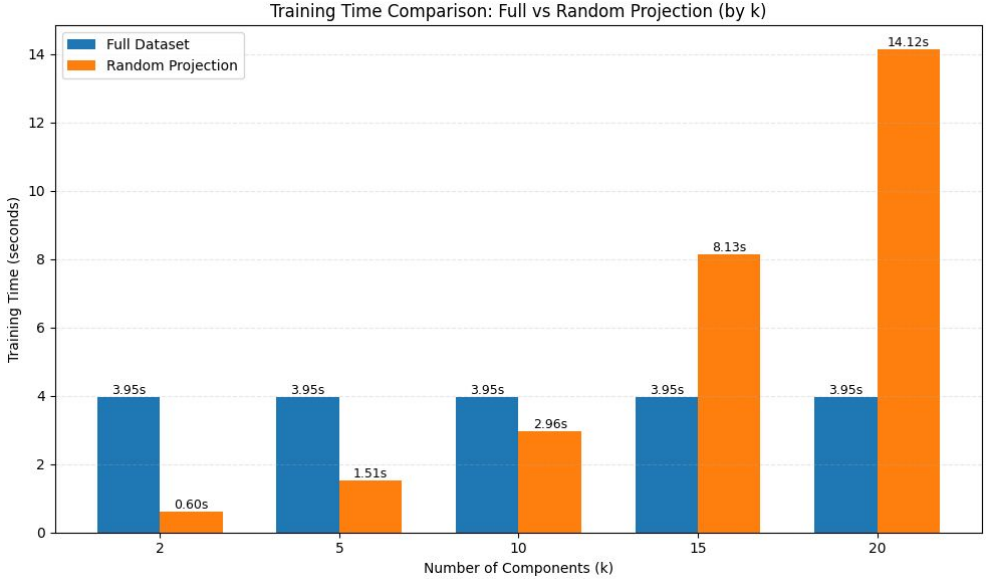
Baseline Model (No Scaling)

- Training complexity: O(nd)
- Test Accuracy: **84.55%**

After Random Projection (Scaling Applied)

- Feature dimension reduced from d to k, where $k \ll d$
- Training complexity becomes: O(nk)
- Test Accuracy: **84.22%**





Training Time Improvement Using Randomized Scaling:-

Conclusion

- Logistic Regression effectively solves the diabetes prediction problem
- Random Projection was used as a randomized scaling technique to speed up training
- Test accuracy without scaling: **84.55%**
- Test accuracy with Random Projection: **84.22%**
- Accuracy drop of only **0.33%**, indicating minimal loss in performance
- Asymptotic training complexity reduced from O(nd) to O(nk)
- asymptotic running time: Logistic Regression ~ O(n * d); after RP it becomes O(n * k) with k << d.
- Demonstrates a favorable trade-off between computational efficiency and accuracy

Future Work

- Experiment with other randomized scaling methods:
  - Johnson–Lindenstrauss Transform
  - Sparse Random Projection
- Apply scaling to more complex models (Support Vector Machine, Neural Networks)
- Evaluate performance on larger and more diverse datasets
- Analyze the effect of different projection dimensions (**k**) more systematically