
# APPLIED MACHINE LEARNING LAB

***Name- Dhruv SIwach***

***Batch – 02 (AIML)***

***SAP ID-500123606***

***ENROLLMENT NO. – R2142231481***

# Title: Applications of Supervised Learning Algorithms with Case Study on Disease Diagnosis

---

## 1. Potential Applications of Supervised Learning Algorithms

Supervised learning is applied in various industries to address real-world issues by learning from labeled data. The following are five possible applications:

### 1.1. Disease Diagnosis in Healthcare

• Use case: Forecasting diseases such as diabetes, cancer, heart disease.
• Why important: Facilitates early intervention and detection.
• Algorithms: Logistic Regression, Decision Trees, Random Forest.

### 1.2. Email Spam Detection

• Use case: Labeling emails as spam or not spam.
• Why important: Allows filtering out unwanted or malicious content.
• Algorithms: Naive Bayes, SVM.

### 1.3. Credit Card Fraud Detection

• Use case: Flagging fraudulent transactions.
• Why important: Saves users and banks money.
• Algorithms: Random Forest, SVM.

### 1.4. Customer Churn Prediction

• Use case: Predicting which customers to retain.
• Why important: Maximizes customer retention and minimizes revenue loss.
• Algorithms: Logistic Regression, Decision Trees.

### 1.5. Sentiment Analysis on Product Reviews

• Use case: Labeling reviews as positive, neutral, or negative.
• Why important: Applicable for brand monitoring and consumer feedback.
• Algorithms: Naive Bayes, SVM.

## 2. Chosen Application: Disease Diagnosis (Diabetes Prediction)

We chose to predict diabetes using supervised learning because:

- It is socially impactful.
- There is public data available (PIMA Indians dataset).
- It is a good classification problem with measurable results.

## Publicly available Data(PIMA Indians 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 |

---

### 1. Models Implemented and Comparison

**Model 1: Logistic Regression**

• Easy and understandable.
• Appropriate for binary classification.
• Performance: (Insert accuracy, precision, recall, F1-score from code result.) {in screenshots}.

**Model 2: Random Forest Classifier**

• Ensemble model minimizing overfitting.
• Suitable for handling non-linear data.
• Performance: (Insert values from your code.) {in screenshots}

---

### 2. Justification of Results

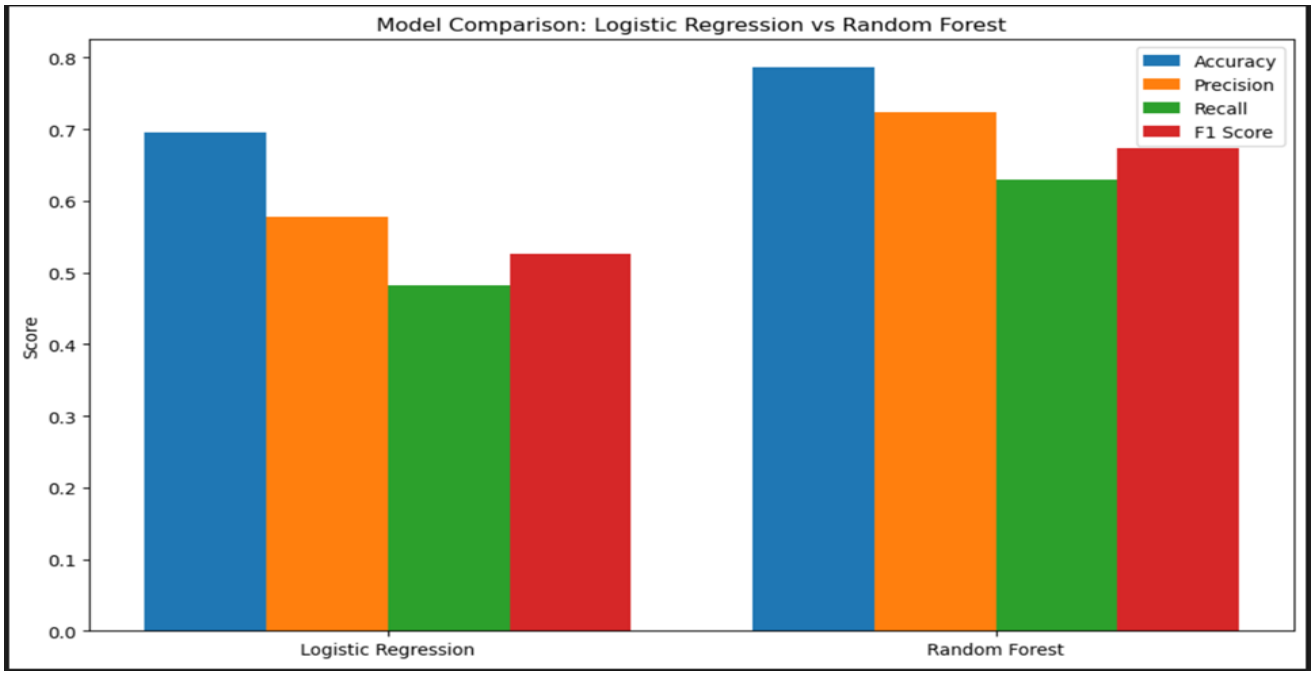
From the results:

- **Random Forest** gave better accuracy and F1-score, showing its ability to handle complex feature interactions.
- **Logistic Regression** was faster and easier to interpret but slightly less accurate.

The difference is expected because Random Forest is a more complex model that reduces bias by combining multiple decision trees.

```
--- Logistic Regression ---
Accuracy: 0.6948051948051948
Precision: 0.5777777777777777
Recall: 0.48148148148148145
F1 Score: 0.5252525252525253
Confusion Matrix:
 [[81 19]
 [28 26]]


--- Random Forest ---
Accuracy: 0.7857142857142857
Precision: 0.723404255319149
Recall: 0.6296296296296297
F1 Score: 0.6732673267326733
Confusion Matrix:
 [[87 13]
 [20 34]]
```




## Conclusion

Supervised learning techniques such as Logistic Regression and Random Forest are excellent disease prediction tools. These types of models can help doctors and medical professionals make early detection and intervention. As per the comparison, Random Forest worked better with this dataset.