KRISH YADAV

Batch – 02 (AIML)

SAP ID-500121939

ENROLLMENT NO. – R2142230521

APPLIED MACHINE LEARNING LAB

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

---

1. Potential Applications of Supervised Learning Algorithms

Supervised learning is used in many industries to solve real-world problems by learning from labelled data. Below are five potential applications:

1.1. Disease Diagnosis in Healthcare

- **Use case**: Predicting diseases like diabetes, cancer, heart disease.
- **Why important**: Enables early detection and intervention.
- **Algorithms**: Logistic Regression, Decision Trees, Random Forest.

1.2. Email Spam Detection

- **Use case**: Classifying emails as spam or not spam.
- **Why important**: Helps filter out unwanted or harmful content.
- **Algorithms**: Naive Bayes, SVM.

1.3. Credit Card Fraud Detection

- **Use case**: Identifying fraudulent transactions.
- **Why important**: Protects users and banks from financial loss.
- **Algorithms**: Random Forest, SVM.

1.4. Customer Churn Prediction

- **Use case**: Predicting which customers will leave a service.
- **Why important**: Helps retain customers and reduce revenue loss.
- **Algorithms**: Logistic Regression, Decision Trees.

1.5. Sentiment Analysis on Product Reviews

- **Use case**: Classifying reviews as positive, neutral, or negative.
- **Why important**: Useful for brand monitoring and 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

---

3. Models Implemented and Comparison

Model 1: Logistic Regression

- Simple and interpretable.
- Works well for binary classification.
- Performance: (Add accuracy, precision, recall, F1-score here from code output.)

Model 2: Random Forest Classifier

- Ensemble model that reduces overfitting.
- Handles non-linear data well.
- Performance: (Add values from your code here.)

---

4. Justification of Results

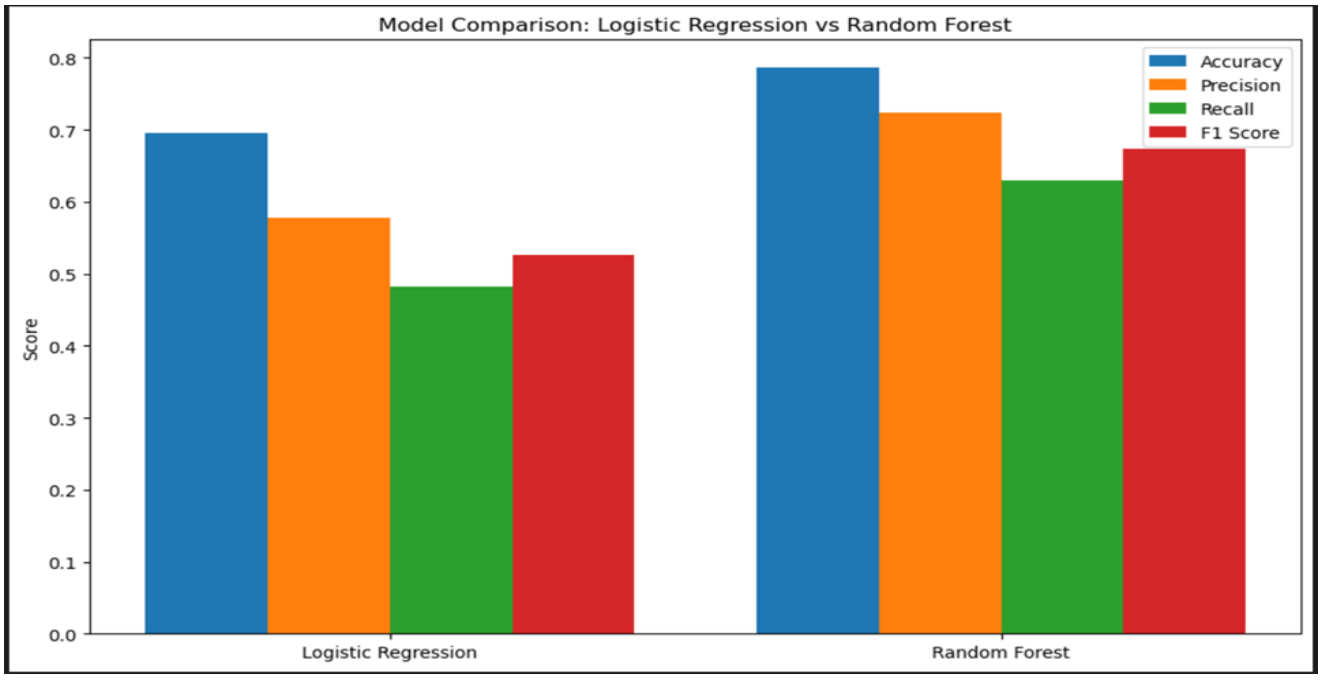
From the results:

- **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 algorithms like Logistic Regression and Random Forest are powerful tools for disease prediction. Such models can assist doctors and health professionals in early detection and intervention. Based on the evaluation, Random Forest performed better for this dataset.