# Diabetes Prediction Report

## Objective:

The primary aim is to develop a machine learning model for predicting diabetes in individuals using the Pima Indians Diabetes dataset. Additionally, an early detection model is implemented to identify potential diabetes cases before the full manifestation of typical symptoms.

## Dataset:

The dataset comprises several medical predictor variables and a target variable (Outcome). Predictors include the number of pregnancies the patient has had, their BMI, insulin level, age, and more. The dataset can be found here: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

## Features:

- Pregnancies

- Glucose

- Blood Pressure

- Skin Thickness

- Insulin

- BMI (Body Mass Index)

- DiabetesPedigreeFunction

- Age

A screenshot of a graph

Description automatically generated

## Target:

- Outcome: Binary classification where 1 denotes the presence of diabetes and 0 denotes its absence.

## Model Development and Results:

## 1. Decision Tree Classifier (with default parameters):

- Training: A Decision Tree classifier was trained using 70% of the data.

- Testing: The model was tested on the remaining 30%.

- Accuracy: The model achieved an accuracy of 70.56%

.

2. Decision Tree Classifier (with entropy criterion & max depth of 3):

- Training: A Decision Tree classifier with entropy as the criterion and a maximum depth of 3 was trained.

- Testing: The model was tested similarly.

3. Visualization: The decision trees were visualized, revealing the decisions made at each node, thus offering insights into how different features contribute to the diabetes prediction.

A diagram of a diagram

Description automatically generated

4. Confusion Matrix: A confusion matrix was plotted for the entropy-based decision tree, showcasing true positives, true negatives, false positives, and false negatives, allowing for a deeper understanding of the model's performance.

A blue squares with white text

Description automatically generated

## 5. Risk Score Generation:

- Using the probabilities provided by the decision tree model, a risk score ranging from 0 to 100 was computed for each individual in the test set, representing their risk of developing diabetes.

Sample 1: Risk Score = 8

Sample 2: Risk Score = 50

Sample 3: Risk Score = 1

Sample 4: Risk Score = 38

Sample 5: Risk Score = 12

Sample 6: Risk Score = 12

Sample 7: Risk Score = 38

Sample 8: Risk Score = 1

Sample 9: Risk Score = 38

Sample 10: Risk Score = 12

Sample 11: Risk Score = 57

Sample 12: Risk Score = 1

Sample 13: Risk Score = 87

Sample 14: Risk Score = 57

Sample 15: Risk Score = 33

Sample 16: Risk Score = 87

Sample 17: Risk Score = 38

Sample 18: Risk Score = 38

Sample 19: Risk Score = 1

Sample 20: Risk Score = 12

Sample 21: Risk Score = 38

Sample 22: Risk Score = 38

Sample 23: Risk Score = 87

Sample 24: Risk Score = 38

Sample 25: Risk Score = 12

Sample 26: Risk Score = 57

Sample 27: Risk Score = 38

Sample 28: Risk Score = 87

Sample 29: Risk Score = 12

Sample 30: Risk Score = 57

Sample 31: Risk Score = 12

Sample 32: Risk Score = 38

Sample 33: Risk Score = 1

Sample 34: Risk Score = 87

Sample 35: Risk Score = 38

Sample 36: Risk Score = 57

Sample 37: Risk Score = 12

Sample 38: Risk Score = 12

Sample 39: Risk Score = 38

Sample 40: Risk Score = 38

Sample 41: Risk Score = 12

Sample 42: Risk Score = 12

Sample 43: Risk Score = 12

Sample 44: Risk Score = 57

Sample 45: Risk Score = 1

Sample 46: Risk Score = 1

Sample 47: Risk Score = 87

Sample 48: Risk Score = 87

Sample 49: Risk Score = 1

Sample 50: Risk Score = 38

Sample 51: Risk Score = 33

Sample 52: Risk Score = 38

Sample 53: Risk Score = 87

Sample 54: Risk Score = 1

Sample 55: Risk Score = 38

Sample 56: Risk Score = 12

Sample 57: Risk Score = 57

Sample 58: Risk Score = 1

Sample 59: Risk Score = 8

Sample 60: Risk Score = 38

Sample 61: Risk Score = 57

Sample 62: Risk Score = 33

Sample 63: Risk Score = 57

Sample 64: Risk Score = 12

Sample 65: Risk Score = 38

Sample 66: Risk Score = 12

Sample 67: Risk Score = 38

Sample 68: Risk Score = 38

Sample 69: Risk Score = 57

Sample 70: Risk Score = 57

Sample 71: Risk Score = 87

Sample 72: Risk Score = 87

Sample 73: Risk Score = 12

Sample 74: Risk Score = 1

Sample 75: Risk Score = 57

Sample 76: Risk Score = 38

Sample 77: Risk Score = 57

Sample 78: Risk Score = 57

Sample 79: Risk Score = 1

Sample 80: Risk Score = 57

Sample 81: Risk Score = 57

Sample 82: Risk Score = 12

Sample 83: Risk Score = 38

Sample 84: Risk Score = 57

Sample 85: Risk Score = 38

Sample 86: Risk Score = 87

Sample 87: Risk Score = 57

Sample 88: Risk Score = 38

Sample 89: Risk Score = 57

Sample 90: Risk Score = 8

Sample 91: Risk Score = 8

Sample 92: Risk Score = 1

Sample 93: Risk Score = 12

Sample 94: Risk Score = 12

Sample 95: Risk Score = 87

Sample 96: Risk Score = 57

Sample 97: Risk Score = 38

Sample 98: Risk Score = 38

Sample 99: Risk Score = 87

Sample 100: Risk Score = 12

Sample 101: Risk Score = 12

Sample 102: Risk Score = 12

Sample 103: Risk Score = 87

Sample 104: Risk Score = 57

Sample 105: Risk Score = 1

Sample 106: Risk Score = 38

Sample 107: Risk Score = 57

Sample 108: Risk Score = 57

Sample 109: Risk Score = 12

Sample 110: Risk Score = 12

Sample 111: Risk Score = 50

Sample 112: Risk Score = 12

Sample 113: Risk Score = 57

Sample 114: Risk Score = 12

Sample 115: Risk Score = 38

Sample 116: Risk Score = 1

Sample 117: Risk Score = 38

Sample 118: Risk Score = 57

Sample 119: Risk Score = 12

Sample 120: Risk Score = 57

Sample 121: Risk Score = 12

Sample 122: Risk Score = 87

Sample 123: Risk Score = 38

Sample 124: Risk Score = 38

Sample 125: Risk Score = 8

Sample 126: Risk Score = 38

Sample 127: Risk Score = 38

Sample 128: Risk Score = 38

Sample 129: Risk Score = 57

Sample 130: Risk Score = 12

Sample 131: Risk Score = 87

Sample 132: Risk Score = 12

Sample 133: Risk Score = 57

Sample 134: Risk Score = 57

Sample 135: Risk Score = 1

Sample 136: Risk Score = 12

Sample 137: Risk Score = 1

Sample 138: Risk Score = 57

Sample 139: Risk Score = 12

Sample 140: Risk Score = 38

Sample 141: Risk Score = 87

Sample 142: Risk Score = 12

Sample 143: Risk Score = 1

Sample 144: Risk Score = 57

Sample 145: Risk Score = 57

Sample 146: Risk Score = 38

Sample 147: Risk Score = 12

Sample 148: Risk Score = 1

Sample 149: Risk Score = 38

Sample 150: Risk Score = 38

Sample 151: Risk Score = 38

Sample 152: Risk Score = 87

Sample 153: Risk Score = 8

Sample 154: Risk Score = 38

Sample 155: Risk Score = 57

Sample 156: Risk Score = 57

Sample 157: Risk Score = 38

Sample 158: Risk Score = 1

Sample 159: Risk Score = 38

Sample 160: Risk Score = 33

Sample 161: Risk Score = 87

Sample 162: Risk Score = 57

Sample 163: Risk Score = 12

Sample 164: Risk Score = 1

Sample 165: Risk Score = 38

Sample 166: Risk Score = 38

Sample 167: Risk Score = 12

Sample 168: Risk Score = 87

Sample 169: Risk Score = 87

Sample 170: Risk Score = 57

Sample 171: Risk Score = 87

Sample 172: Risk Score = 87

Sample 173: Risk Score = 12

Sample 174: Risk Score = 12

Sample 175: Risk Score = 87

Sample 176: Risk Score = 38

Sample 177: Risk Score = 50

Sample 178: Risk Score = 12

Sample 179: Risk Score = 1

Sample 180: Risk Score = 38

Sample 181: Risk Score = 38

Sample 182: Risk Score = 87

Sample 183: Risk Score = 12

Sample 184: Risk Score = 87

Sample 185: Risk Score = 1

Sample 186: Risk Score = 38

Sample 187: Risk Score = 1

Sample 188: Risk Score = 87

Sample 189: Risk Score = 57

Sample 190: Risk Score = 38

Sample 191: Risk Score = 38

Sample 192: Risk Score = 1

Sample 193: Risk Score = 12

Sample 194: Risk Score = 1

Sample 195: Risk Score = 87

Sample 196: Risk Score = 1

Sample 197: Risk Score = 38

Sample 198: Risk Score = 12

Sample 199: Risk Score = 12

Sample 200: Risk Score = 87

Sample 201: Risk Score = 87

Sample 202: Risk Score = 57

Sample 203: Risk Score = 38

Sample 204: Risk Score = 12

Sample 205: Risk Score = 50

Sample 206: Risk Score = 87

Sample 207: Risk Score = 1

Sample 208: Risk Score = 38

Sample 209: Risk Score = 12

Sample 210: Risk Score = 57

Sample 211: Risk Score = 12

Sample 212: Risk Score = 57

Sample 213: Risk Score = 38

Sample 214: Risk Score = 50

Sample 215: Risk Score = 12

Sample 216: Risk Score = 57

Sample 217: Risk Score = 38

Sample 218: Risk Score = 12

Sample 219: Risk Score = 12

Sample 220: Risk Score = 57

Sample 221: Risk Score = 1

Sample 222: Risk Score = 1

Sample 223: Risk Score = 12

Sample 224: Risk Score = 87

Sample 225: Risk Score = 8

Sample 226: Risk Score = 57

Sample 227: Risk Score = 12

Sample 228: Risk Score = 1

Sample 229: Risk Score = 1

Sample 230: Risk Score = 87

Sample 231: Risk Score = 38

## 6. Early Detection Model:

- Hypothesis: Glucose level above 120 is indicative of early diabetes.

- Oversampling: Due to the imbalance in early diabetes detection, the training set was oversampled using SMOTE to balance the classes.

- Model: A Decision Tree classifier was used to identify early cases of diabetes.

- Result: The early detection model could predict diabetes in its preliminary stages, even before all typical symptoms are evident.

array([1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0,

1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,

0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1,

0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,

1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0,

0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,

0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1,

1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,

0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,

0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0])

## Recommendations and Conclusion:

1. The decision tree classifier, particularly the one with entropy as a criterion, provides decent accuracy and is interpretable, making it useful for medical professionals to understand the basis of predictions.

2. The risk score is a helpful tool for patients and healthcare providers to understand the relative risk of diabetes for an individual and act accordingly.

3. The early detection model serves as a promising tool for proactive healthcare, enabling interventions even before diabetes becomes full-blown.

For a more effective prediction and early detection, it is advisable to further explore different machine learning algorithms, feature engineering, and possibly include more relevant features, if available.