Modou K Touray

3734028

Methods for selecting the features subsets

- **Information Gain (Mutual Information)**: Measures how much information each feature provides about the target variable. Features with high mutual information have strong predictive power.

- **Gain Ratio**: Adjusts Information Gain by taking into account the intrinsic information of each feature, reducing the bias toward features with many unique values.

- **Correlation-Based Selection**: Measures the correlation between each feature and the target variable. Features that have high correlation with the target are likely to be predictive.

Classifier Details and Test Results on Information Gain Subset

Metric	Decision Tree	Naive Bayes	Random Forest
Accuracy	**0.99**	**0.24**	**0.99**
Precision	**0.97**	**0.22**	**1.00**
Recall	**0.96**	**0.98**	**0.95**
TP	**411**	**418**	**406**
FP	**13**	**1485**	**0**
TN	**1529**	**57**	**1542**
FN	**16**	**9**	**21**

- **Decision Tree** and **Random Forest** classifiers perform well, with high accuracy, precision, and recall.
- The **Random Forest classifier** is the best choice due to its perfect precision for fraudulent cases and no false positives, making it highly reliable in minimizing false alerts.
- **Naive Bayes** performs poorly, likely due to the strong independence assumption, which doesn't fit well with this dataset's feature correlations.

Classifier Details and Test Results on Gain Ratio Subset

Metric	Decision Tree	Naive Bayes	Random Forest
Accuracy	**0.98**	**0.23**	**0.99**
Precision	**0.95**	**0.22**	**1.00**
Recall	**0.96**	**0.99**	**0.95**
TP	**409**	**423**	**406**
FP	**22**	**1519**	**0**
TN	**1520**	**23**	**1542**
FN	**18**	**4**	**21**

- **Random Forest** is the best-performing classifier, achieving the highest precision with no false positives and an excellent overall accuracy of 99%.
- **Decision Tree** also performs well, but it has a slightly lower precision and a few more false positives compared to Random Forest.
- **Naive Bayes** performs poorly, with a very low accuracy of 23%, largely due to the strong independence assumption, which may not fit this dataset.

Classifier Details and Test Results on Correlation Subset

Metric	Decision Tree	Naive Bayes	Random Forest
Accuracy	**0.99**	**0.40**	**0.99**
Precision	**0.98**	**0.27**	**0.98**
Recall	**0.97**	**0.99**	**0.97**
TP	**415**	**422**	**415**
FP	**10**	**1169**	**8**
TN	**1532**	**373**	**1534**
FN	**12**	**5**	**12**

- Both Decision Tree and Random Forest classifiers perform excellently on the Correlation subset, with high accuracy, precision, and recall, making them reliable choices for fraud detection.
- Random Forest slightly outperforms Decision Tree by having a lower number of false positives, resulting in fewer incorrect fraud alerts.
- Naive Bayes performs poorly with a very low accuracy of 40% and a high false positive rate, suggesting it is not suitable for this dataset due to its independence assumption.

Overall Best Model and Subset

Best Model: The **Random Forest classifier** consistently outperforms the other models across all subsets with high accuracy, precision, and recall.

Best Feature Subset:

- The **Information Gain** and **Gain Ratio** subsets both provide excellent results for the Random Forest classifier, with a perfect precision of 100% and accuracy of 99%.
- **Correlation subset** also performs well, but the Information Gain and Gain Ratio subsets have a slight edge due to their consistently high precision and balanced performance across metrics.

Conclusion

- **Best Model**: **Random Forest** (high precision and recall, minimal false positives).
- **Best Subset**: **Information Gain** and seconded by **Gain Ratio** for optimal results with Random Forest.