

Results

From the model performance table, the top three models are as follows:

- 1. **Random Forest Model with Robust and Min-Max Normalization:** Achieved the best performance with 100% accuracy and a Root Mean Squared Error (RMSE) of 0. This model significantly exceeded the assignment's minimum accuracy requirement of 75%.
- 2. **Random Forest Model with Z-Score Normalization:** Delivered exceptional results with an accuracy of 99.9994%, maintaining the model's consistency across different normalization techniques.
- 3. **Decision Tree Model with Robust Normalization:** Also achieved 100% accuracy, matching the top Random Forest models in terms of accuracy, though it lacked the same level of consistency across other techniques.

Data Tabulation

Table 1. Data Properties 1: Types of Data and Data Types

Variable	Types of Data	Data Types	Measurement Level	Range	Min Value	Top Value	Unique Values	Null Values	Outliers
Distance from Home	Numeric	float64	Ratio	0.004874385 - 10632.72367	0.004874385	10632.72367	999950	0	Yes
Distance from Last Transaction	Numeric	float64	Ratio	0.000118282 - 11851.10456	0.000118282	11851.10456	999799	0	Yes
Ratio to Median Purchase Price	Numeric	float64	Ratio	0.00439924 - 267.8029422	0.00439924	267.8029422	999775	0	Yes
Repeat Retailer	Numeric	float64	Ratio	0.0 - 1.0	0.0	1.0	2	0	No
Used Chip	Numeric	float64	Ratio	0.0 - 1.0	0.0	1.0	2	0	No
Used Pin Number	Numeric	float64	Ratio	0.0 - 1.0	0.0	1.0	2	0	No
Online Order	Numeric	float64	Ratio	0.0 - 1.0	0.0	1.0	2	0	No
fraud	Numeric	float64	Ratio	0.0 - 1.0	0.0	1.0	2	0	Yes

Table 2. Data Properties 2: Statistics

Variable	Frequency	Percentile (25th, 50th, 75th)	Data Completeness	Mean	Median	Mode	Standard Deviation	Variance	Skewness	Kurtosis
Distance from Home	-	3.877956, 9.967709, 25.743846	100.0	26.628778897080572	9.9677089035	1.202045073	65.39116353562537	4276.0042685429	20.239662692496626	1471.6021122705229
Distance from Last Transaction	-	0.296643, 0.998626, 3.355676	100.0	5.036496408638979	0.9986257605	0.010964433	25.843271311862804	667.874672098551	125.92038849698808	46978.63510152133
Ratio to Median Purchase Price	-	0.475633, 0.997692, 2.09636	100.0	1.8241549873166671	0.9976924645	0.108386876	2.7995918747827666	7.837714665349686	8.91508784502601	289.5063272919147
Repeat Retailer	-	1.0, 1.0, 1.0	100.0	0.8815383415367816	1.0	1.0	0.32315414026037587	0.10442859836742267	-2.3613443684692115	3.575954378506143
Used Chip	-	0.0, 0.0, 1.0	100.0	0.350395055428776	0.0	0.0	0.4770941652793858	0.22701884254363388	0.627148284585248	-1.6864882425672337
Used Pin Number	-	0.0, 0.0, 0.0	100.0	0.10060640848971886	0.0	0.0	0.3008069971703368	0.09048484954663502	2.6554860363443056	5.051616192589408
Online Order	-	0.0, 1.0, 1.0	100.0	0.6505571077995091	1.0	1.0	0.4767942791466928	0.22733278462701442	-0.6315403330257717	-1.601160010130552
fraud	-	0.0, 0.0, 0.0	100.0	0.08740422365913122	0.0	0.0	0.28242663668924295	0.07976480511159763	2.9217966575837475	6.536908782264169

Conclusion

We selected the **Random Forest model** and applied **Min-Max normalization** to the dataset. With optimized hyperparameters (hyp), the model achieved impressive performance, attaining an accuracy of (acc) of 100% accuracy and a Root Mean Squared Error (rmse) of 0.