(RED)*

WINE

MACHINE LEARNING

02

CAT & TIM

DATA SELECTION & PREPARATION

- Red variants of Portuguese wine. Source: Kaggle
  - 1599 rows, 12 columns

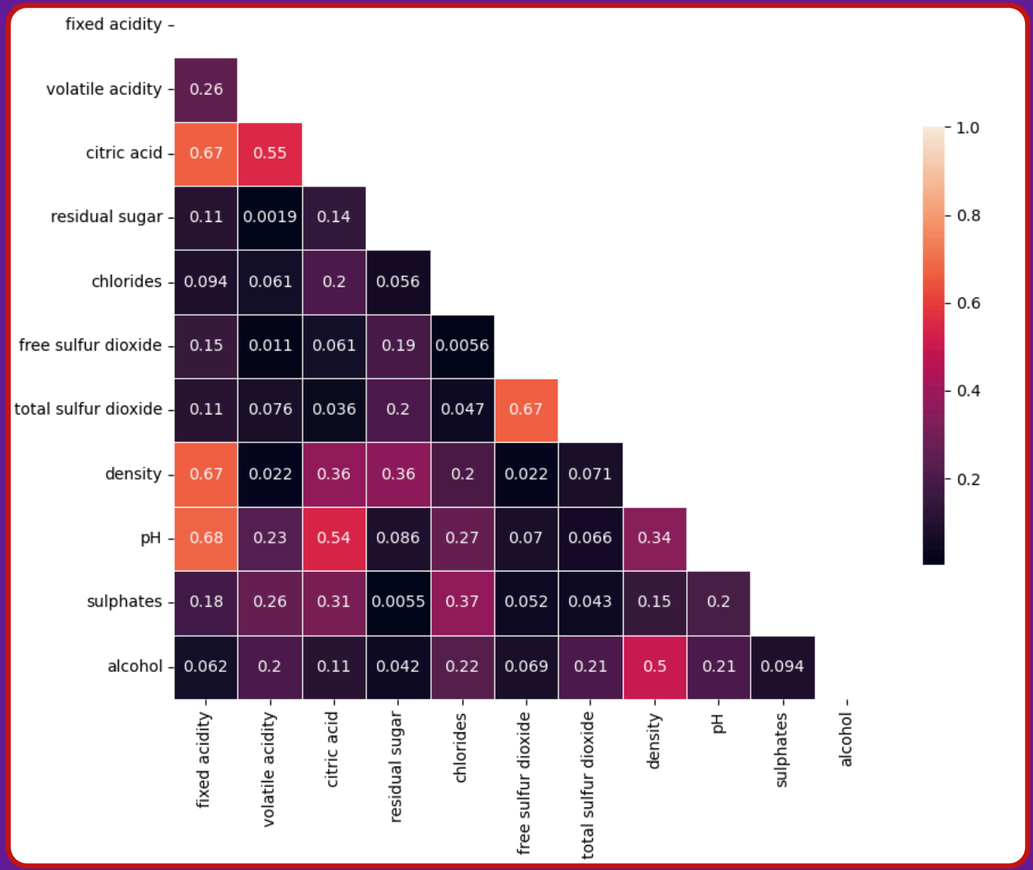
- Data Preparation:
  - Checked for null values
  - Removed existing Quality column
  - Created Target Column ‘Good Quality’

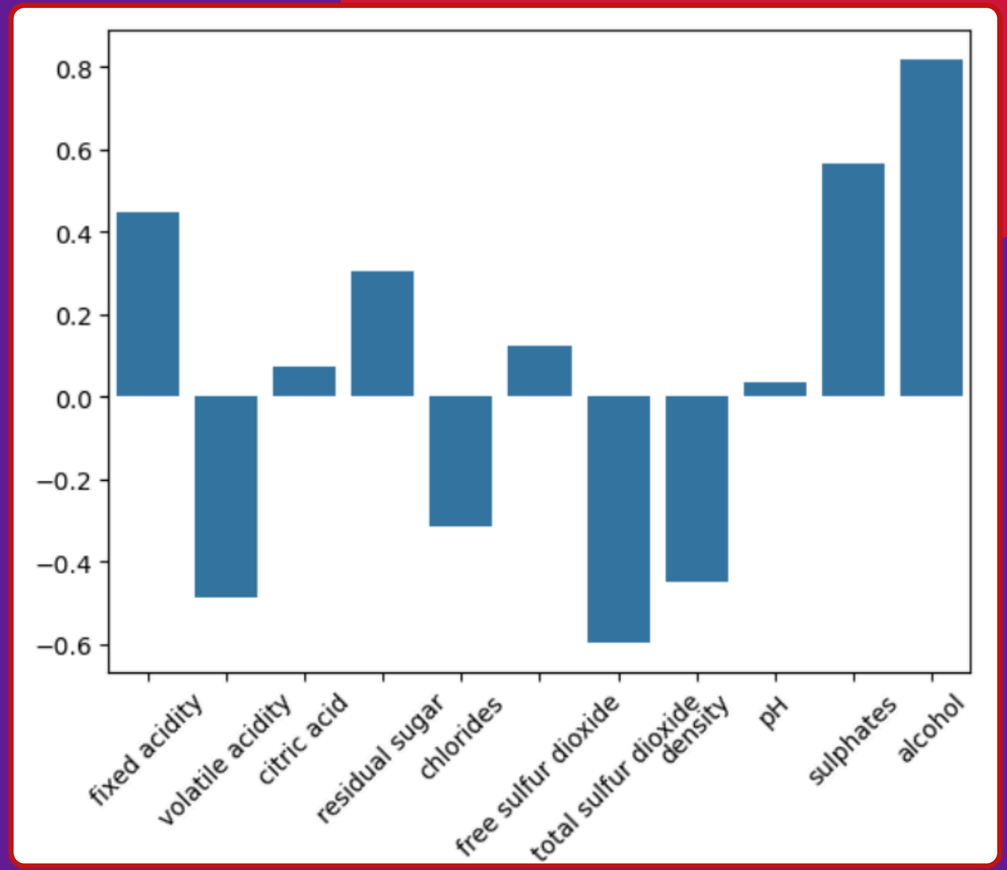
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

FEATURE ENGINEERING & SELECTION

- Co-efficiency Barplot
- Correlation Matrix
- Normalizer: MinMaxScaler, Standard Scaler
- Handling Imbalances

CORRELATION





MODEL BUILDING & EVALUATION

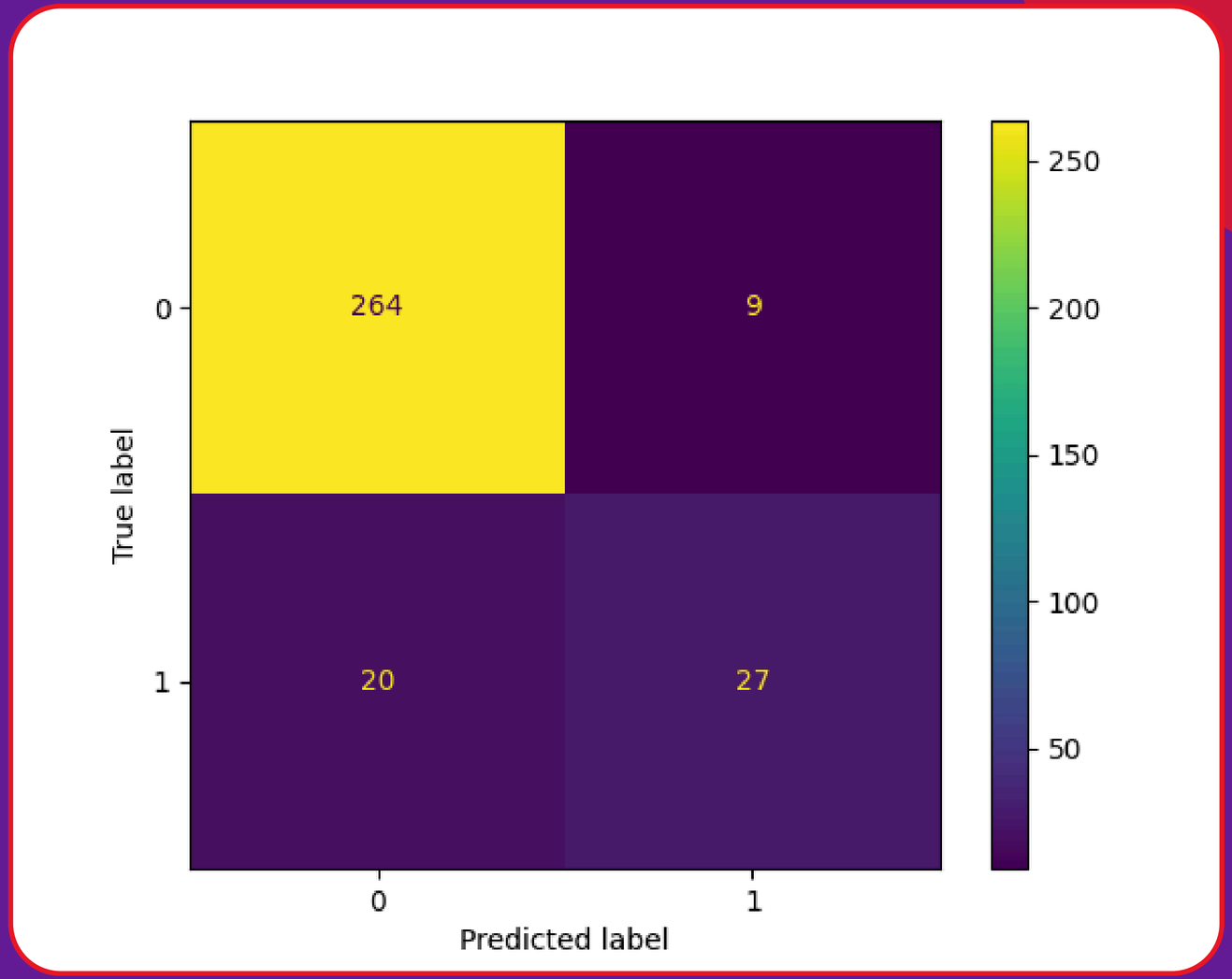
	Precision	Recall	Accuracy
KNN*	0.48	0.34	0.85
Logistic Regression	0.25	0.36	0.75
Decision Tree**	0.63	0.36	0.88

**KNN: 3 Neighbors
**Decision tree: depth 5

ENSEMBLES

	Precision	Recall	Accuracy
Bagging	0.62	0.45	0.88
Random Forest	0.74	0.55	0.91
Gradient Boosting	0.60	0.55	0.88
Ada Boost	0.63	0.62	0.89

CONFUSION MATRIX



MODEL OPTIMIZATION

Problem

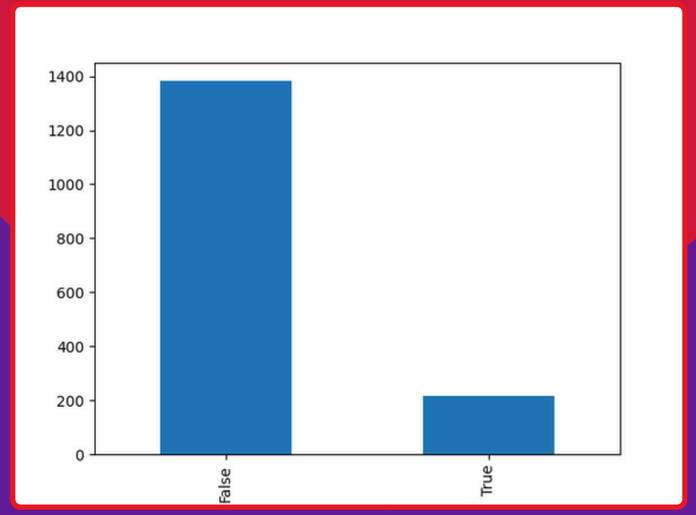
- Imbalanced data - more bad than good wine

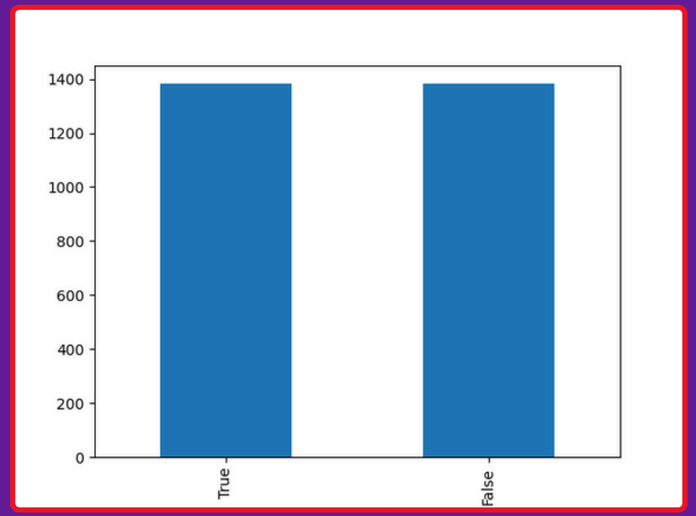
Approach

- Oversampling to balance the dataset
- Resampled good wines to match number of bad
- Train RandomForestClassifier on balanced data

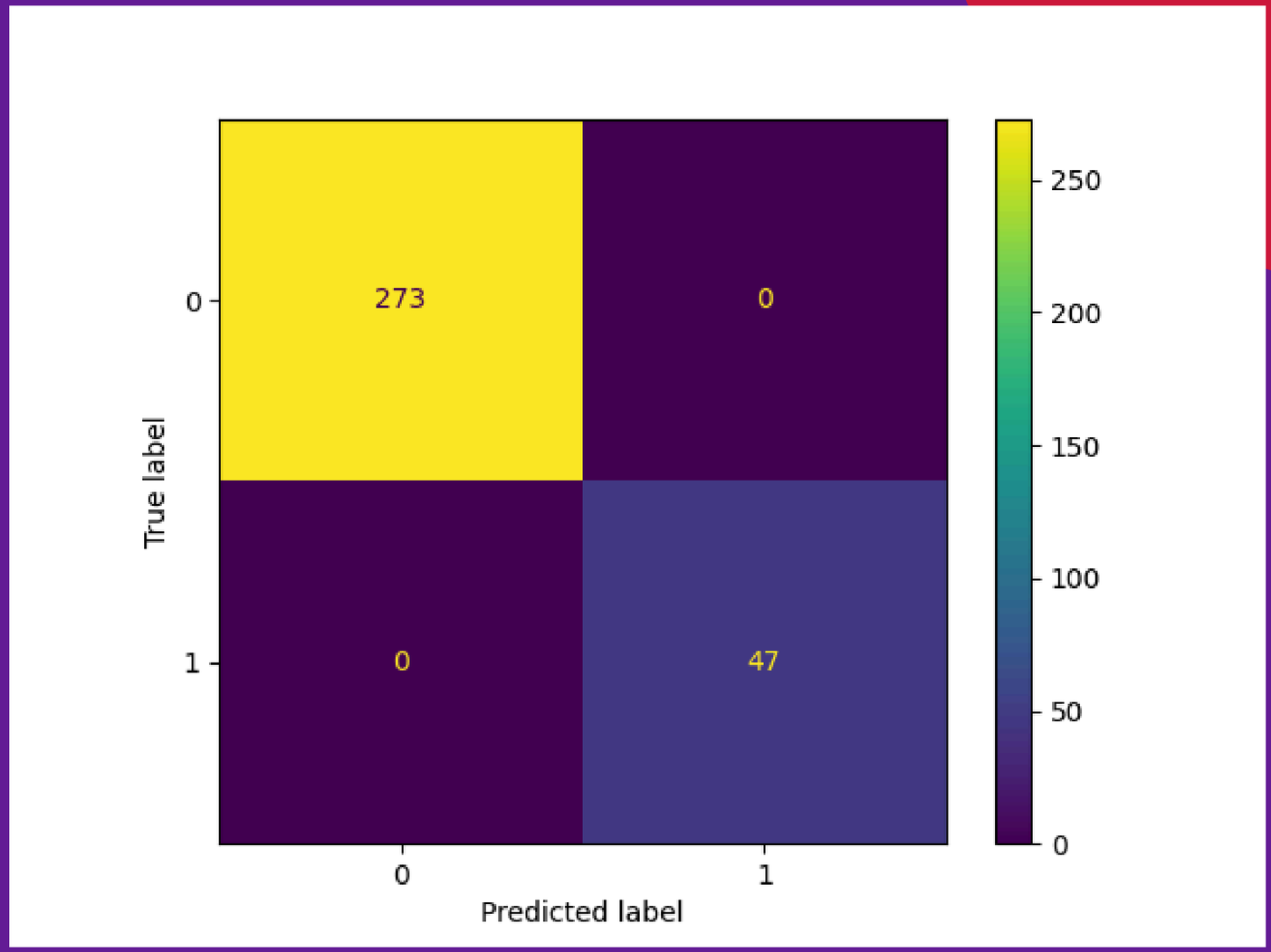
Result

- Before: 91%; After: 100%
- Overfitting? -> More validation





CONFUSION MATRIX



CONCLUSION

- RandomForestClassifier proved best in order to eliminate False Positives and increase precision rate
- After getting rid of imbalances results turned out to be overfitted

REAL WORLD APPLICATION

- Helps winemakers optimize quality control and improve consistency
- Faster assessments, cost savings, and better customer satisfaction

- Bias Risk: Model learns from historical data.
- Overfitting: 100% accuracy may not generalize well.

CHALLANGES & FUTURE

Challenges, Findings

- Selecting the right model, Imbalanced data
- Precision and Recall matter
- Perfect score can signal overfitting

Future Work, Improvements

- Include sensory data and environmental factors
- Bias evaluation in training data

THANK YOU