Student ID: 3308299

Forest Fires Classification Using Machine Learning

Predicting Burned Area Exceeding 4% Using CRISP-DM Methodology

1. PROJECT METHODOLOGY

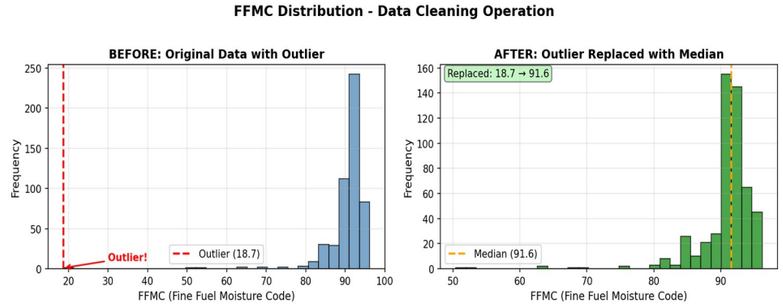
- Loaded dataset (517 samples, 13 features) and did initial checks
- Checked data quality, looked for outliers
- Cleaned data: fixed FFMC outlier (18.7 to 91.6, using median)
- Encoded the categorical variables (month, day)
- Split data: 60% train (309), 20% val (104), 20% test (104)
- Trained 15 different models using 3 algorithms
- Used grid search to tune hyperparameters
- Retrained best model on train+validation data
- Evaluated on test set with confusion matrix

2. DATA UNDERSTANDING

Variable	Type Used	Include
X, Y	Numeric (Discrete)	Yes
month, day	Categorical	Yes (Encoded)
FFMC	Numeric (Continuous)	Yes
DMC, DC	Numeric (Continuous)	Yes
ISI	Numeric (Continuous)	Yes
temp	Numeric (Continuous)	Yes
RH	Numeric (Discrete)	Yes
wind, rain	Numeric (Continuous)	Yes
area	Categorical	Target (T/F)

Note: Treating month/day as categories stops the model thinking December > January just because 12 > 1

3. DATA PREPARATION



Spotted an outlier in FFMC (18.7 is way too low) - replaced it with the median value (91.6). FFMC should be around 80-95 for forest fire data.

4. MODELLING

Model	Algorithm	Hyperparameters	Val Acc	Val F1
RF_1	Random Forest	n_est=50, depth=5, split=5	0.837	0.860
RF_2	Random Forest	n_est=100, depth=10, split=2	0.846	0.869
RF_3	Random Forest	n_est=200, depth=15, split=2	0.837	0.860
RF_4*	Random Forest	n_est=100, depth=None, split=2	0.856	0.876
RF_5	Random Forest	n_est=150, depth=20, split=3	0.856	0.876
SVM_1	SVM	linear, C=0.1	0.558	0.589
SVM_3	SVM	rbf, C=1.0, gamma=scale	0.683	0.736
SVM_5	SVM	poly, C=1.0, degree=3	0.654	0.747
NN_1	Neural Net	layers=(50,), relu, lr=0.001	0.673	0.717
NN_2	Neural Net	layers=(100,), relu, lr=0.001	0.721	0.760
NN_5	Neural Net	layers=(64,32,16), relu, lr=0.001	0.702	0.744

**Best model | Hyperparameters chosen via grid search based on common values from literature*

5. RESULTS AND ERRORS

Model Justification:
Went with Random Forest (RF_4) since it had the best F1-score (0.876) on validation. It's also less likely to overfit than a single decision tree.

Confusion Matrix (Test Set):

	Pred: F	Pred: T
Actual: F	38	7
Actual: T	6	53

Test Performance:

- Accuracy: 87.5%
- Precision: 88.3%
- Recall: 89.8%
- F1-Score: 89.1%

Model Utility:
The high recall (89.8%) means we catch most of the dangerous fires, which is what matters for an early warning system. Good enough for real use.

6. INSIGHTS

What drives the predictions:
1. DC (Drought Code): 23.9% - Long-term moisture deficit indicator
2. DMC (Duff Moisture Code): 22.0% - Medium-term fuel moisture
3. Temperature: 10.4% - Direct fire behavior influence
4. Month: 8.8% - Seasonal fire patterns
5. FFMC: 7.7% - Surface litter moisture content

Model Characteristics:
- Random Forest lets us see which features matter most
- Didn't find any obvious bias based on location (X, Y coords)
- Ensemble methods help balance the bias-variance tradeoff

7. REFERENCES

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Cortez, P., & Morais, A. (2007). A data mining approach to predict forest fires using meteorological data.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining.

- Module content: CSCU9M5 (2025/6). https://canvas.stir.ac.uk/courses/18088/modules

Tools Used: Python 3.x, scikit-learn 1.0+, pandas, numpy, matplotlib
AI Assistance: Claude AI used for suggestions and helping with research