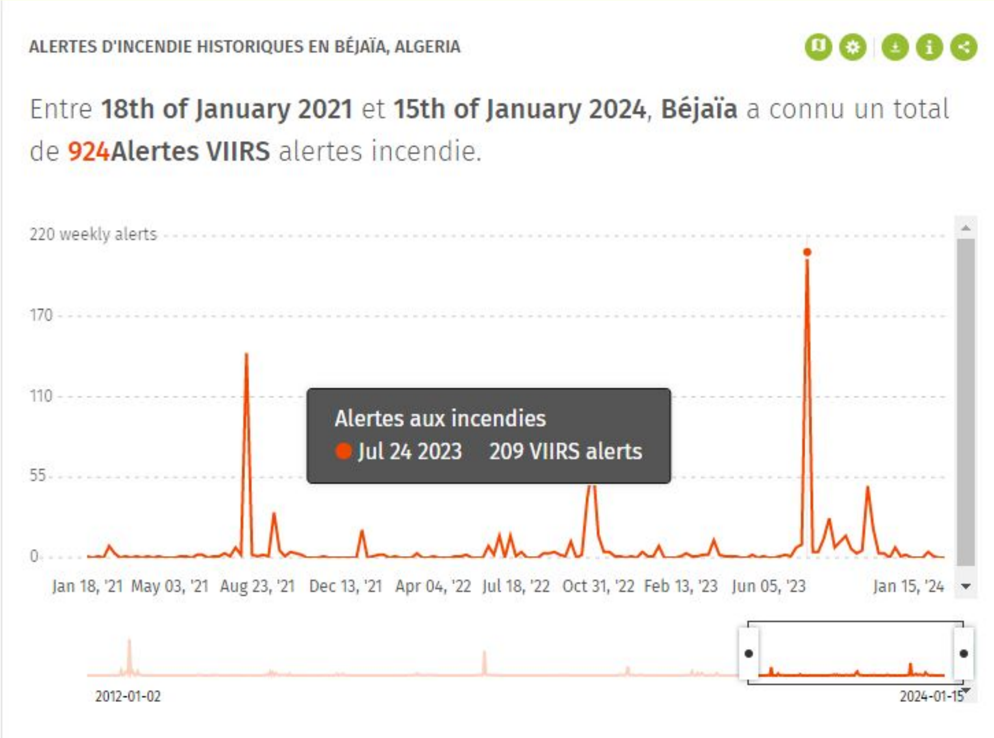
Predicting Wildfires in Bejaia

Data Mining Project Presentation

Introduction

Introduction



Introduction

Due to the intensity of these fires, The development of a predictive model for early detection and intervention can provide crucial insights to authorities.

Dataset

Dataset #1

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
0	1	6	2012	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	not fire
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0	not fire
4	5	6	2012	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	not fire
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
239	26	9	2013	30	65	14	0.0	85.4	16.0	44.5	4.5	16.9	6.5	fire
240	27	9	2013	28	87	15	4.4	41.1	6.5	8	0.1	6.2	0	not fire
241	28	9	2013	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	not fire
242	29	9	2013	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	not fire
243	30	9	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	not fire

244 rows × 14 columns

Dataset

Dataset #2

	date	Temperature	Rain	Wd	Ws	Pres	RH	dew point Max	dew point Avg	dew point Min	Classes
0	2023-06-01	22.3	0.2	158.0	9.2	1013.7	79.6	17.0	15.5	11.0	0
1	2023-06-02	23.4	0.0	157.0	9.1	1012.4	81.2	18.0	15.9	14.0	0
2	2023-06-03	20.8	0.0	237.0	8.7	1014.3	83.4	17.0	15.9	14.0	0
3	2023-06-04	24.0	9.6	270.0	8.7	1013.4	79.8	18.0	16.6	15.0	0
4	2023-06-05	24.3	0.0	97.0	9.1	1014.4	80.0	19.0	16.4	9.0	0
...	...	...	...	...	...	...	...	...	...	...	...
1769	2012-09-26	31.0	0.0	NaN	11.0	1009.4	54.0	23.0	20.3	17.0	not fire
1770	2012-09-27	31.0	0.0	NaN	11.0	1010.0	66.0	23.0	21.0	19.0	fire
1771	2012-09-28	32.0	0.7	NaN	14.0	1007.3	47.0	24.0	21.5	20.0	not fire
1772	2012-09-29	26.0	1.8	NaN	16.0	1012.8	80.0	20.0	18.3	16.0	not fire
1773	2012-09-30	25.0	1.4	NaN	14.0	1015.7	78.0	19.0	17.3	15.0	not fire

1774 rows × 11 columns

Appendix

- Data Pre-Processing
  - Data Exploration
  - Data Cleaning
  - Data Visualization & interpretation

- Data Dimensionality Reduction
  - Supervised & Unsupervised Feature Selection
  - PCA
- Model Implementation
  - Decision Tree
  - Applying different models (Logistic Regression, KNN, Decision Tree & Random Forest)
- Conclusion

Data Pre-Processing

Data exploration , Data Cleaning, Handling Outliers..etc

Data Exploration

Information about data

DataSet #1

```
[ ]  # general information about dataset
     data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   day          244 non-null    int64
 1   month        244 non-null    int64
 2   year         244 non-null    int64
 3   Temperature  244 non-null    int64
 4    RH          244 non-null    int64
 5    Ws          244 non-null    int64
 6   Rain         244 non-null    float64
 7   FFMC         244 non-null    float64
 8   DMC          244 non-null    float64
 9   DC           244 non-null    object
 10  ISI          244 non-null    float64
 11  BUI          244 non-null    float64
 12  FWI          244 non-null    object
 13  Classes      243 non-null    object
dtypes: float64(5), int64(6), object(3)
memory usage: 26.8+ KB
```

Dataset #2

```
[ ]  # general information about dataset
     data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1774 entries, 0 to 1773
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   date            1774 non-null   object
 1   Temperature     1620 non-null   float64
 2   Rain            1756 non-null   float64
 3   Wd              839 non-null    float64
 4   Ws              1523 non-null   float64
 5   Pres            1041 non-null   float64
 6   RH              1768 non-null   float64
 7   dew point Max   1768 non-null   float64
 8   dew point Avg   1768 non-null   float64
 9   dew point Min   1768 non-null   float64
 10  Classes         1773 non-null   object
dtypes: float64(9), object(2)
memory usage: 152.6+ KB
```

Data Exploration

The Null Values we Have found

DataSet #1

```
day             0
month           0
year            0
Temperature     0
 RH             0
 Ws             0
Rain            0
FFMC            0
DMC             0
DC              0
ISI             0
BUI             0
FWI             0
Classes         1
dtype: int64
```

Dataset #2

```
date                  0
Temperature         154
Rain                 18
Wd                  935
Ws                  251
Pres                733
RH                    6
dew point Max         6
dew point Avg         6
dew point Min         6
Classes               1
dtype: int64
```

Data Exploration

Summary

DataSet #1

- -The dataset contains 14 columns (included indexes) and 244 rows
- -Only one row is containing a null value on the target attributes, its position 165
- -Some statistical measures show unexpected values due to the datatype of attributes.

Dataset #2

- The dataset contains 11 columns and 1774 rows
- The datatype of two columns ('date' and 'Classes') attributes is object, but the good thing is that the rest of the columns are in their appropriate datatypes.
- There is a considerable amount of null values.
- Some statistical measures show unexpected values..

We will carry on the presentation with the perspective of the First DATASET.

Almost same steps are followed for the 2nd dataset

Data Cleaning

Summary

- Fixed attribute names (UpperCase)
- Fixed attribute Data Types
- Fixed shifted data
- Unify the values (eg: not fire ,fire)
- Remove white spaces
- Remove duplicates
- Explore outliers

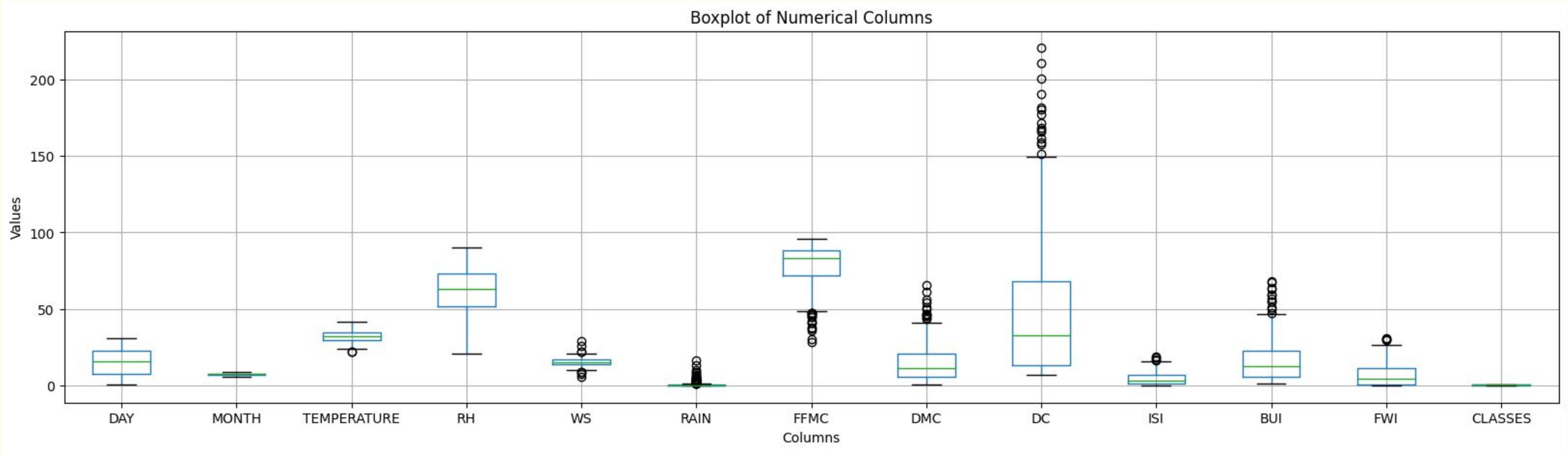
```
#make all columns name upperCases
columns_Formatted = [col.upper() for col in data.columns ]
#take the new columns names
data.columns = columns_Formatted
data.columns
```

```
Index(['DAY', 'MONTH', 'YEAR', 'TEMPERATURE', 'RH', 'WS', 'RAIN', 'FFMC',
       'DMC', 'DC', 'ISI', 'BUI', 'FWI', 'CLASSES'],
      dtype='object')
```

	DAY	MONTH	YEAR	TEMPERATURE	RH	WS	RAIN	FFMC	DMC	DC	ISI	BUI	FWI	CLASSES
165	14	7	2013	37.0	37.0	18.0	0.2	88.9	12.9	14.6 9	12.5	10.4	fire	NaN

Data Visualization and Interpretation

Summary

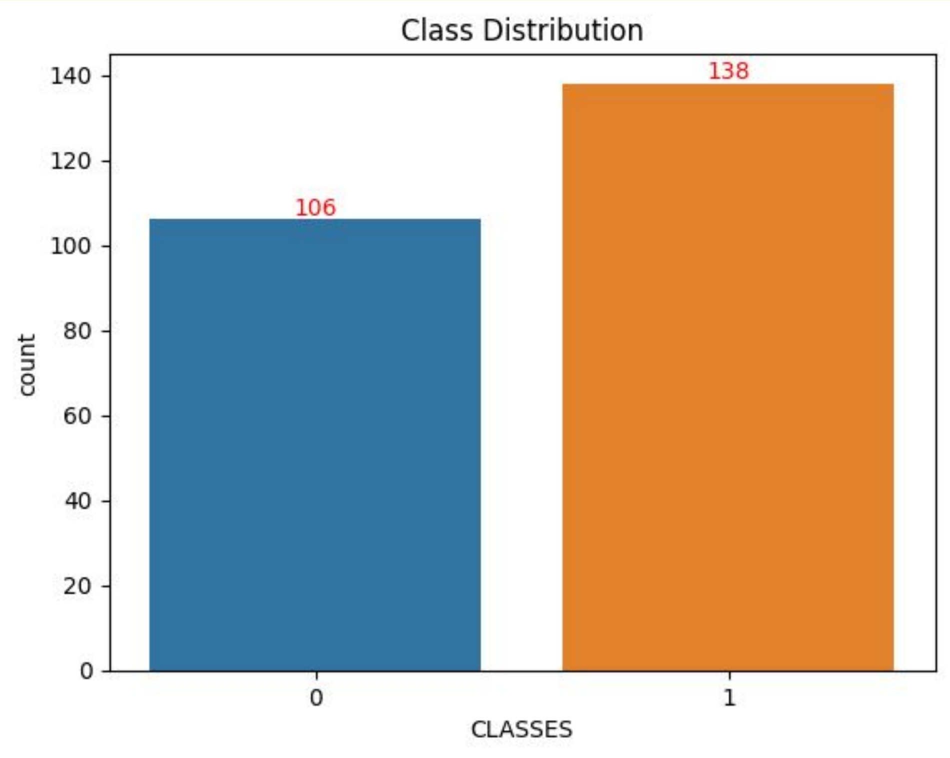


a definitive decision on whether to eliminate or adjust these outliers will be determined after implementing scaling methods, such as Z-score analysis.

Data Visualization and Interpretation

Summary

- The bar chart shows there are more instances of fire *138* than non-fire events *106*.
- This difference is essential to note for creating accurate models.
- When training our prediction system, we need to be mindful of this imbalance to make sure our model doesn't get skewed towards one of them.



Data Visualization and Interpretation

Summary

- We also added box plots that illustrates distribution of weather variables (Bui,RH..) compared to classes (0:not fire 1:fire)

- + Lineplots for the evolution of those weather variables (RH,Temperature..) according to time (Month).

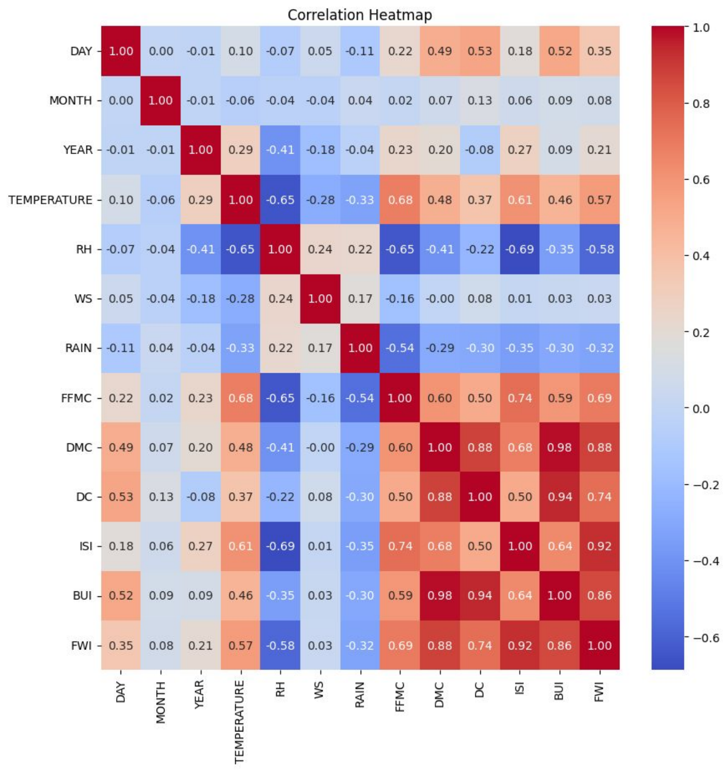
Data Dimensionality Reduction

- Supervised & Unsupervised Feature Selection
- PCA

Unsupervised Feature Selection

Summary

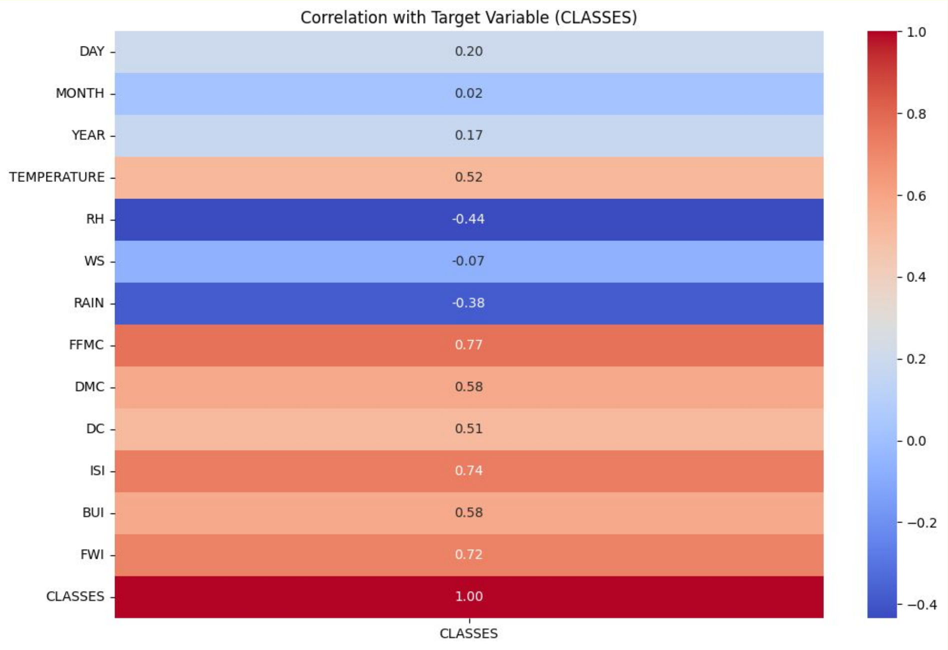
- The Correlation Heatmap reveals strong positive correlations among the components within the Fire Weather Index (FWI) system, suggesting potential *redundancy*.
- Data is too similar, keeping all these similar attributes can confuse our analysis.



Supervised Feature Selection

Summary

- Strong positive correlations are observed with FFMC, ISI, and FWI, emphasizing their significant influence on fire events.
- These insights guide the identification of crucial factors contributing to wildfires in Bejaia, guiding the development of our predictive model.



Forward Selection

Summary

The forward selection algorithm identified 'RH', 'FFMC', 'ISI', and 'FWI' as the best set of features for a classification problem related to fires.

Relative Humidity, Fine Fuel Moisture Code, Initial Spread Index

And Fire Weather Index.

```
#the result of the forward selection
best_att
```

```
Index(['RH', 'FFMC', 'ISI', 'FWI'], dtype='object')
```

PCA

Summary

	PC_1	PC_2	PC_3	PC_4	PC_5	CLASSES
0	-47.680279	3.406184	-6.552777	-12.169857	-1.738406	0
1	-47.904387	-0.085685	-5.718294	-10.743381	-0.228223	0
2	-53.700134	-26.603496	-8.557722	-10.122543	6.474630	0
3	-57.763164	-41.623813	-17.983954	-11.271860	8.277542	0
4	-43.170082	-13.904976	3.222945	-7.036598	1.457301	0

Model Implementation

- Decision Tree
- Applying different models (Logistic Regression, KNN, Decision Tree & Random Forest)

Decision Tree

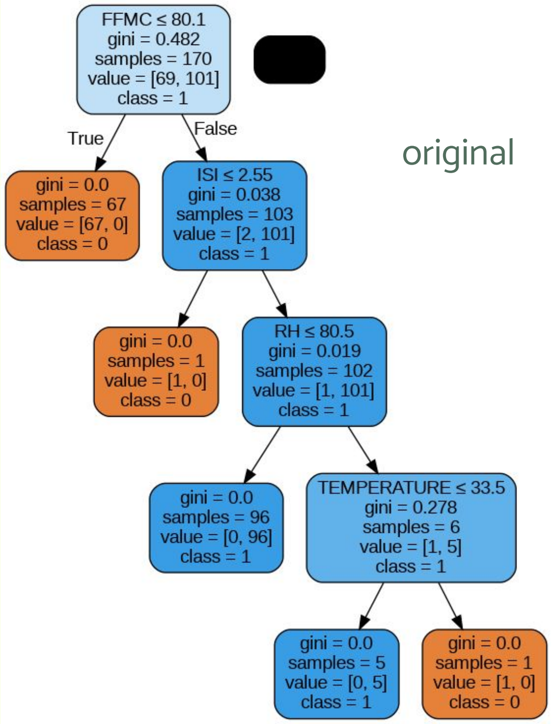
Accuracy

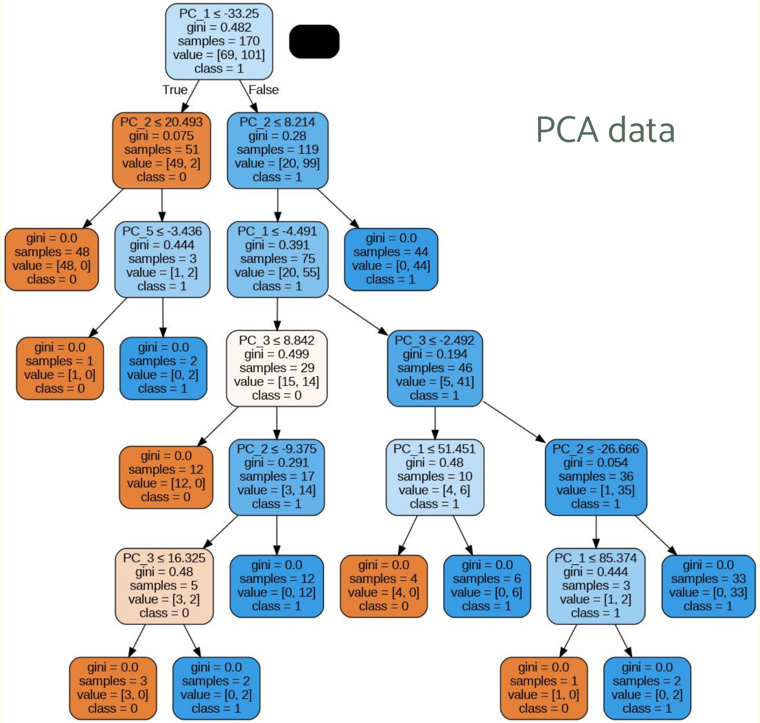
- The model performs exceptionally well across both classes, achieving high precision (0.97, 1), recall (1, 0.97), and F1-score (0.99, 0.99), indicating strong predictive ability.
- With an accuracy of 98% to 99%, it shows an excellent performance in predicting the target values.
- The result of Model from the Original data seem much better than the data generated from PCA
- To dive deeper and optimize the model's performance while considering computational efficiency, we aim to identify the best parameter values applicable to the decision tree algorithm

```
Model Accuracy for orgignal data
Accuracy: 0.972972972972973
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        37
           1       0.97      0.97      0.97        37

    accuracy                           0.97        74
   macro avg       0.97      0.97      0.97        74
weighted avg       0.97      0.97      0.97        74

--------------------------------------------
Model Accuracy for PCA generated data
Accuracy: 0.8918918918918919
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.91      0.89        35
           1       0.92      0.87      0.89        39

    accuracy                           0.89        74
   macro avg       0.89      0.89      0.89        74
weighted avg       0.89      0.89      0.89        74

--------------------------------------------
```

Decision Tree

non parametric decision tree using holdout method on : Original data, PCA generated data, T-SNE generated data and data after features selection

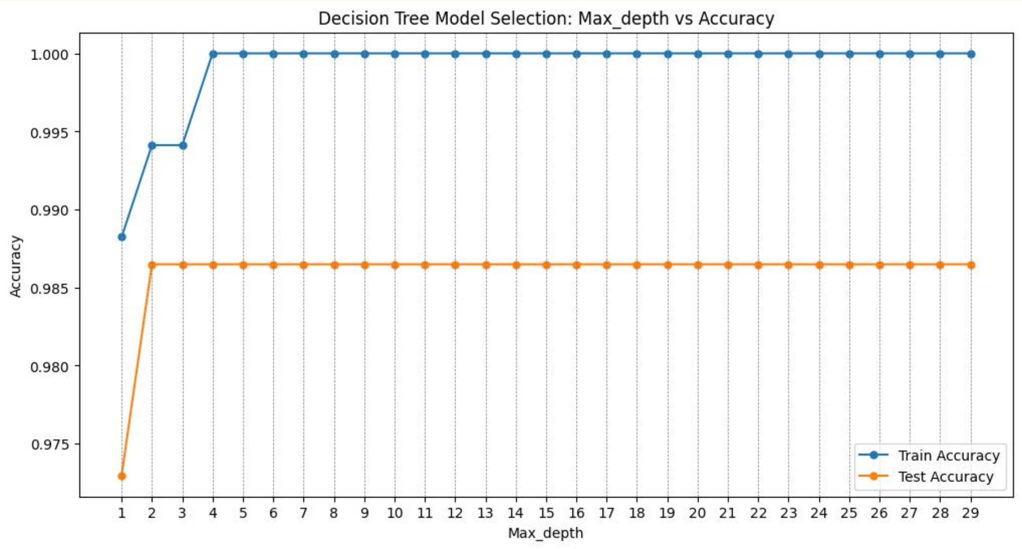




Decision Tree

Second Try: we will use parametric decision tree using holdout method and find the best prameters

- As we can see through this graph representation the optimal value of min samples leaf is 2 which leads to very excellent accuarcy 1 and for max depth the optimal value is above or equal 2 wihch leads to accuracy above 0.985
- In general this model preform very good in predection the target values with different values of parameters



Applying different Models

Logistic Regression, KNN, Decision Tree & Random Forest

Other Techniques

- **Holdout Method**: Dividing the dataset into training and validation sets.
- **Cross-Validation**: Employing cross-validation to robustly assess model performance.
- **GridSearch Optimization**: Utilizing GridSearch to optimize hyperparameters for enhanced model performance.

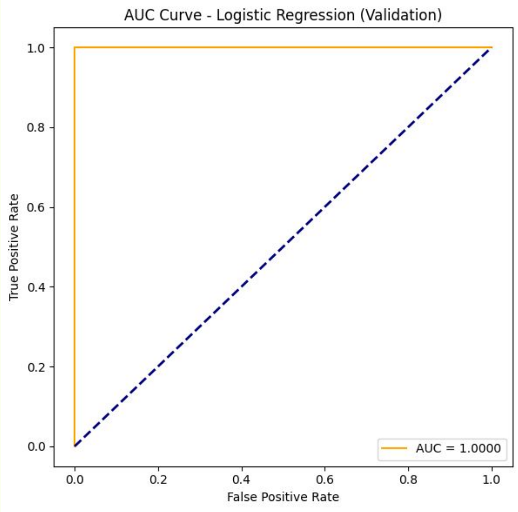
Evaluation Metrics

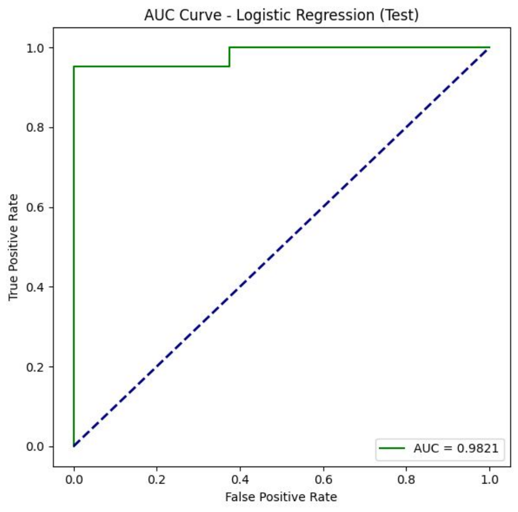
- **Accuracy**: Determining the ratio of correctly predicted instances to the total instances.
- **F-score**: Balancing precision and recall for assessing model accuracy.
- **Precision**: Evaluating the ratio of correctly predicted positive observations to the total predicted positives.
- **Recall**: Assessing the ratio of correctly predicted positive observations to the all actual positives.

Applying different Models

Logistic Regression, KNN, Decision Tree & Random Forest

- Some plots :






```
# Create a dictionary of models
models = {
    'Logistic Regression': LogisticRegression(),
    'KNN': KNeighborsClassifier(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier()
}
```

Applying different Models

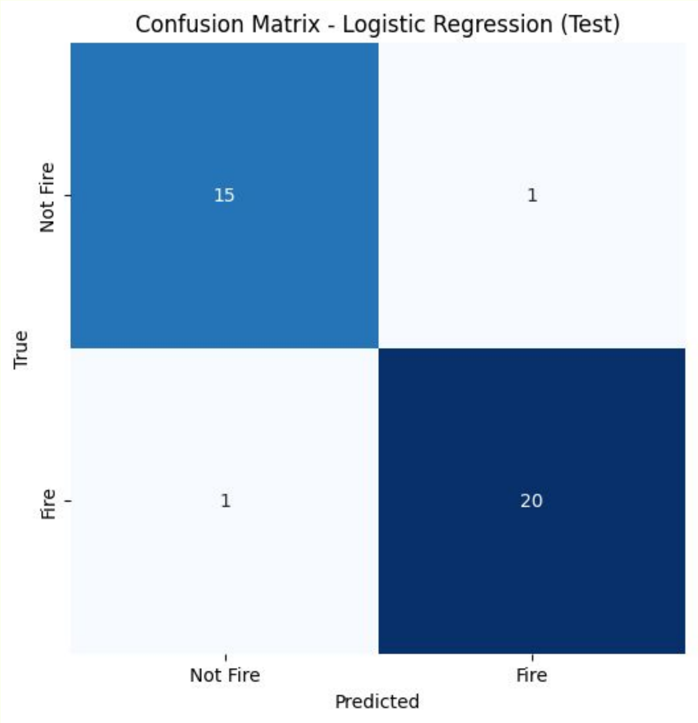
Logistic Regression

Logistic Regression:

Best Parameters: {'C': 100} Best Score: 0.9538 Validation Accuracy: 0.9388 Validation F1 Score: 0.9434 Validation Recall: 0.9259 Validation Precision: 0.9615

Interpretation:

- Cross-validation helps mitigate overfitting concerns.
- The high precision of 96.15% indicates that when the model predicts the presence of fire, it is correct 96.15% of the time.
- The model can be considered suitable for predicting fires in Bejaia, providing a balanced trade-off between precision and recall.



Applying different Models

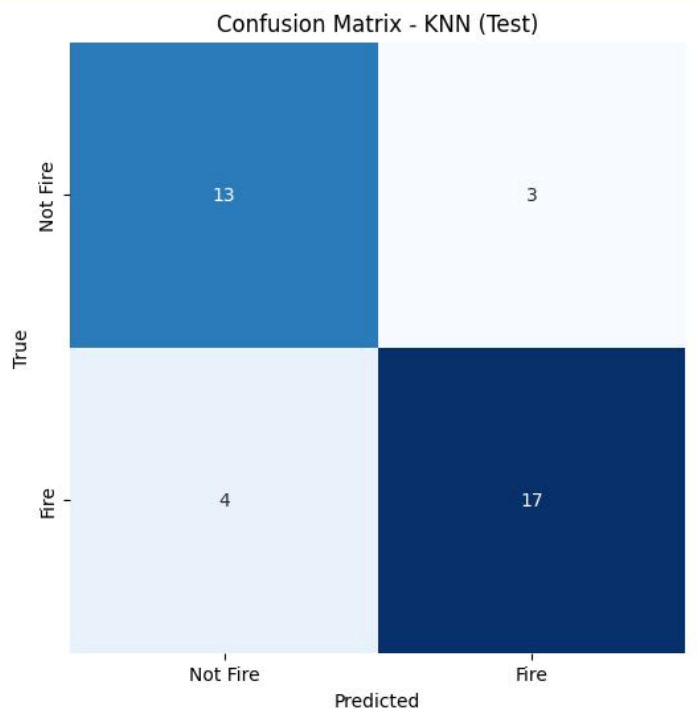
KNN

KNN:

Best Parameters: {'n_neighbors': 9} Best Score: 0.8462 Validation Accuracy: 0.8571 Validation F1 Score: 0.8727 Validation Recall: 0.8889 Validation Precision: 0.8571

Interpretation:

- KNN provides a decent performance but appears to lag behind logistic regression in terms of cross-validated accuracy (84.62%).
- While cross-validation helps mitigate overfitting, it suggests that the model may not generalize as well as logistic regression.
- The accuracy, F1 score, recall, and precision are reasonable but not as high as logistic regression.



Applying different Models

Random Forest

Random Forest:

Best Parameters: {'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 50} Best Score: 0.9795 Validation Accuracy: 1.0000 Validation F1 Score: 1.0000 Validation Recall: 1.0000 Validation Precision: 1.0000

Interpretation:

Random Forest, like the Decision Tree, achieves perfect scores in both training and cross-validation. The complexity introduced by the ensemble approach may contribute to overfitting concerns. Further investigation is needed to ensure that the model generalizes well to new data, especially considering the small dataset size.

Regarding 2nd Dataset

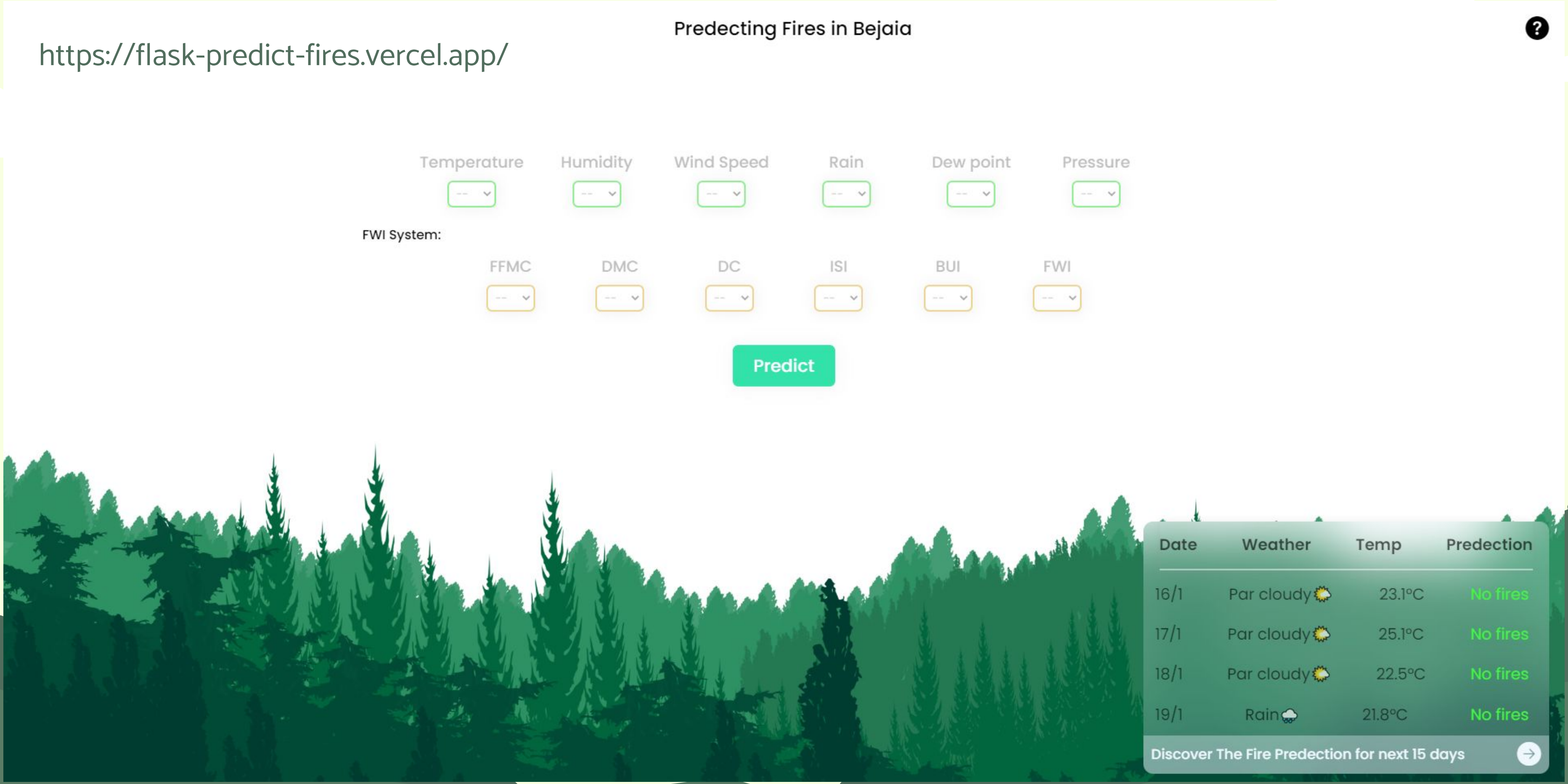
Surprise

3

2

1🎉

https://flask-predict-fires.vercel.app/



Predicting Fires in Bejaia

Date	Weather	Temp	Rain	Humidity	Wind speed	Dew point	Prediction
16/1	Par cloudy	23.1°C	0.0	47.1%	14.9km	5.9°C	No fires
17/1	Par cloudy	25.1°C	0.0	29.7%	15.0km	1.4°C	No fires
18/1	Par cloudy	22.5°C	0.0	47.1%	12.8km	6.2°C	No fires
19/1	Rain	21.8°C	0.4	59.5%	15.7km	9.2°C	No fires
20/1	Rain	15.7°C	5.7	70.3%	13.9km	8.3°C	No fires
21/1	Rain	11.8°C	2.4	79.3%	7.4km	7.3°C	No fires
22/1	Rain	14.0°C	0.0	69.2%	6.9km	4.2°C	No fires
23/1	Rain	15.3°C	0.1	64.4%	6.9km	4.5°C	No fires
24/1	Sunny	13.9°C	0.0	72.7%	5.6km	5.6°C	No fires
25/1	Par cloudy	17.7°C	0.0	48.4%	6.0km	2.0°C	No fires
26/1	Par cloudy	17.4°C	0.0	45.1%	6.3km	1.3°C	No fires
27/1	Sunny	16.3°C	0.0	52.9%	5.8km	3.2°C	No fires
28/1	Par cloudy	16.2°C	0.0	66.9%	5.4km	6.3°C	No fires
29/1	Sunny	18.7°C	0.0	57.9%	7.2km	5.4°C	No fires
30/1	Par cloudy	15.7°C	0.0	69.1%	5.6km	7.2°C	No fires

Conclusion

- The deployment of predictive models for forecasting wildfires in the Bejaia region of Algeria carries significant real-world implications where they offer a reliable tool for early detection, allowing authorities to take swift actions in mitigating fire risks.
-The positive impact extends to minimizing property damage, safeguarding lives, and optimizing resource allocation.
- Moreover, the use of such models enables a proactive approach to firefighting efforts, enabling authorities to respond promptly to identified high-risk areas.
- Beyond immediate applications, the development and enhancement of these models, incorporating advanced techniques like deep learning, hold promise for further refinement.