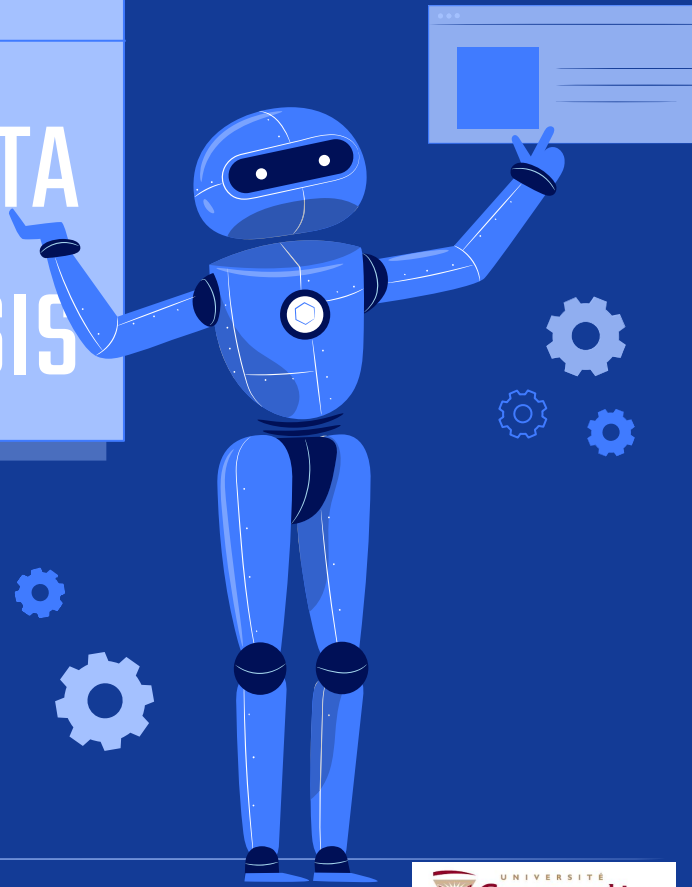
US WILDFIRES DATA ANALYSIS

Analysis of 24 years of geo-referenced wildfire records.

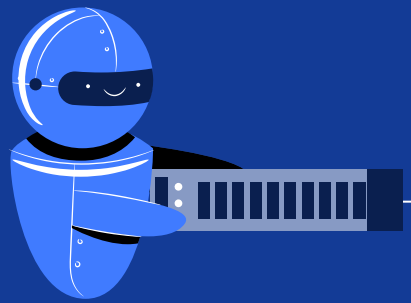


Presented by:
Arvind, Darshan, Yogesh and Sourabh

Université Concordia University

TABLE OF CONTENTS

01 INTRODUCTION



PROBLEM STATEMENT

01.

8.8 MILLION acres burned in 2018.

02.

$3.1 BILLION overall firefighting costs in 2018

OBJECTIVES

Fire size prediction

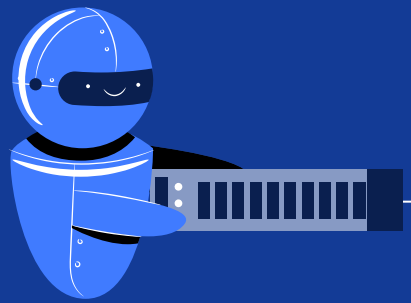
Clustering based on wildfire attributes

Fire Cause prediction

RELATED WORK

- Analysis of Machine Learning Methods for Wildfire Security Monitoring with an Unmanned Aerial Vehicles
- Machine learning to predict final fire size at the time of ignition-International Journal of Wildland Fire(https://www.publish.csiro.au/WF/WF19023)-Dmitriy Alexandrov, Elizaveta Pertseva, Ivan Berman, Igor Pantiukhin, Aleksandr Kapitonov
- Data-driven Forest Fire analysis, Jerry Gao, Kshama Shalini, Navit Gaur, Xuan Guan

02 METHODOLOGY



DATASET

- Source:Kaggle
- Features:
  - Real Valued and String
  - Feature Count : 39
  - The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations.
  - The dataset, referred to as the Fire Program Analysis fire-occurrence database (FPA FOD), includes 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned from 1992 to 2015.

DATASET

	OBJECTID	FOD ID	FPA ID	SOURCE S	SOURCE SY	NWCG RE	NWCG REI	NWCG REPORTING UNIT N	SOURCE F	SOURCE REPORTING UN	LOCAL FI	LOCAL	FIRE COD	FIRE NAME	ICS 209 IN	ICS 209	MTBS ID	MTBS FIR	COMF	FIRE
	1	1	FS-1418826	FED	FS-FIRESTAT	FS	USCAPNF	Plumas National Forest	0511	Plumas National Forest	1	PNF-47	BJ8K	FOUNTAIN	NULL	NULL	NULL	NULL	NULL	2005
	2	2	FS-1418827	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	13	13	AAC0	PIGEON	NULL	NULL	NULL	NULL	NULL	2004
	3	3	FS-1418835	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	27	021	A32W	SLACK	NULL	NULL	NULL	NULL	NULL	2004
	4	4	FS-1418845	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	43	6	NULL	DEER	NULL	NULL	NULL	NULL	NULL	2004
	5	5	FS-1418847	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	44	7	NULL	STEVENOT	NULL	NULL	NULL	NULL	NULL	2004
	6	6	FS-1418849	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	54	8	NULL	HIDDEN	NULL	NULL	NULL	NULL	NULL	2004
	7	7	FS-1418851	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	58	9	NULL	FORK	NULL	NULL	NULL	NULL	NULL	2004
	8	8	FS-1418854	FED	FS-FIRESTAT	FS	USCASHF	Shasta-Trinity National Fo...	0514	Shasta-Trinity National ...	3	02	BK5X	SLATE	NULL	NULL	NULL	NULL	NULL	2005
	9	9	FS-1418856	FED	FS-FIRESTAT	FS	USCASHF	Shasta-Trinity National Fo...	0514	Shasta-Trinity National ...	5	03	BLPQ	SHASTA	NULL	NULL	NULL	NULL	NULL	2005
0	10	10	FS-1418859	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	61	10	NULL	TANGLEF...	NULL	NULL	NULL	NULL	NULL	2004
1	11	11	FS-1418861	FED	FS-FIRESTAT	FS	USCAENF	Eldorado National Forest	0503	Eldorado National Forest	64	11	NULL	FORK #2	NULL	NULL	NULL	NULL	NULL	2004

TECHNOLOGIES

PySpark

https://spark.apache.org/docs/latest/api/python/pyspark.html

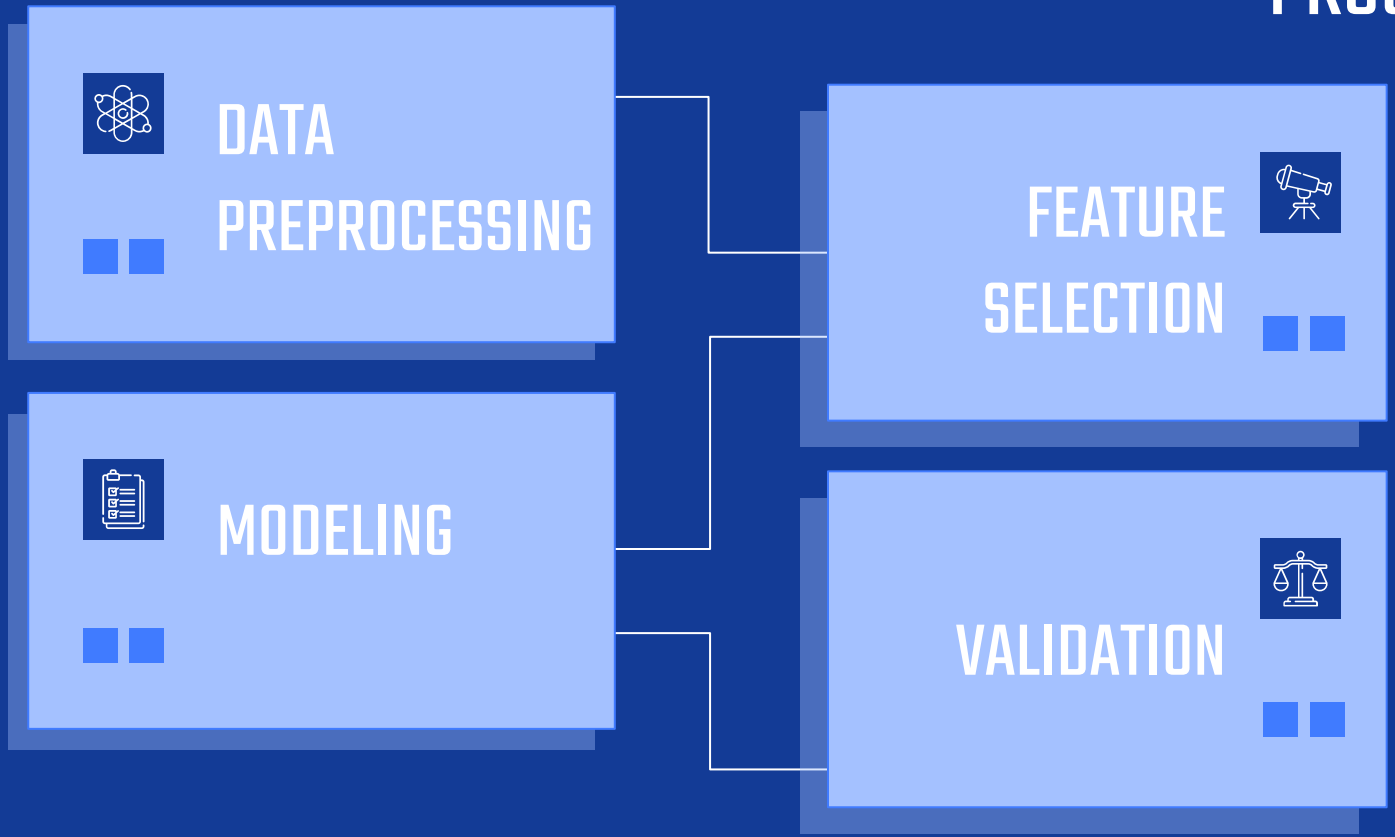
scikit-learn

https://scikit-learn.org/stable/

matplotlib

https://matplotlib.org/

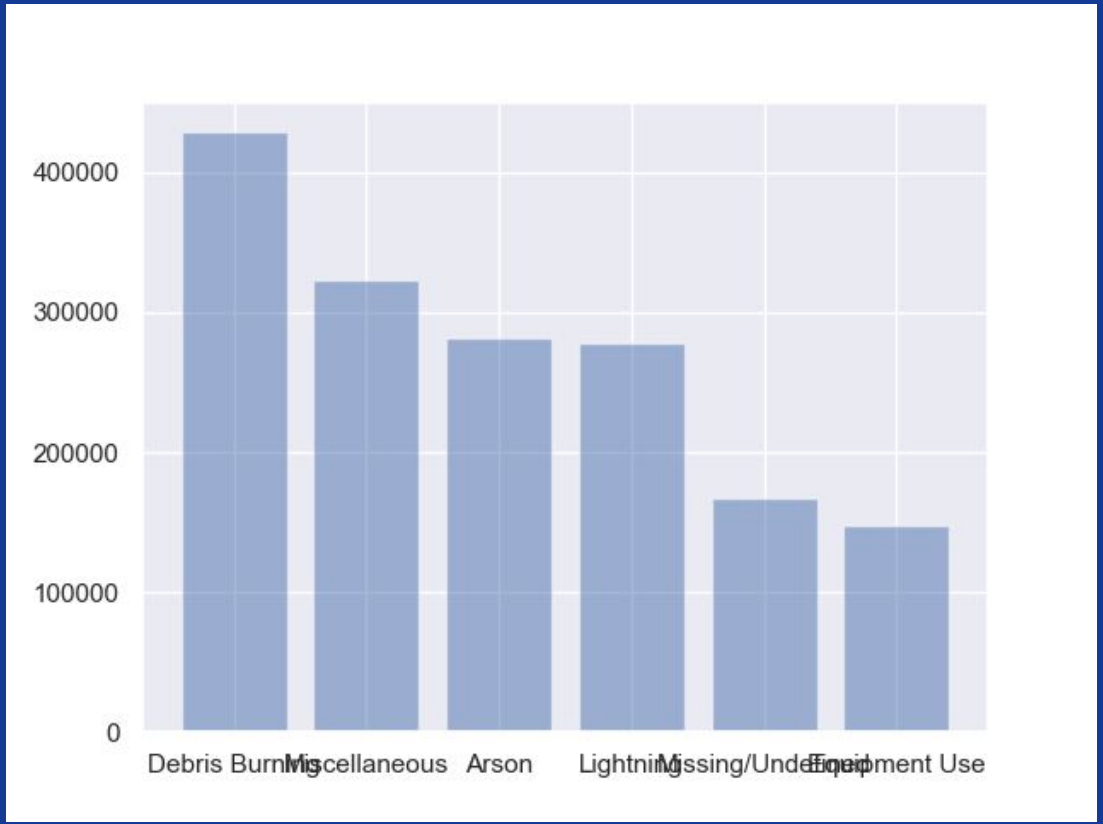
PROCESS


DATA PREPROCESSING
MODELING
FEATURE SELECTION
VALIDATION

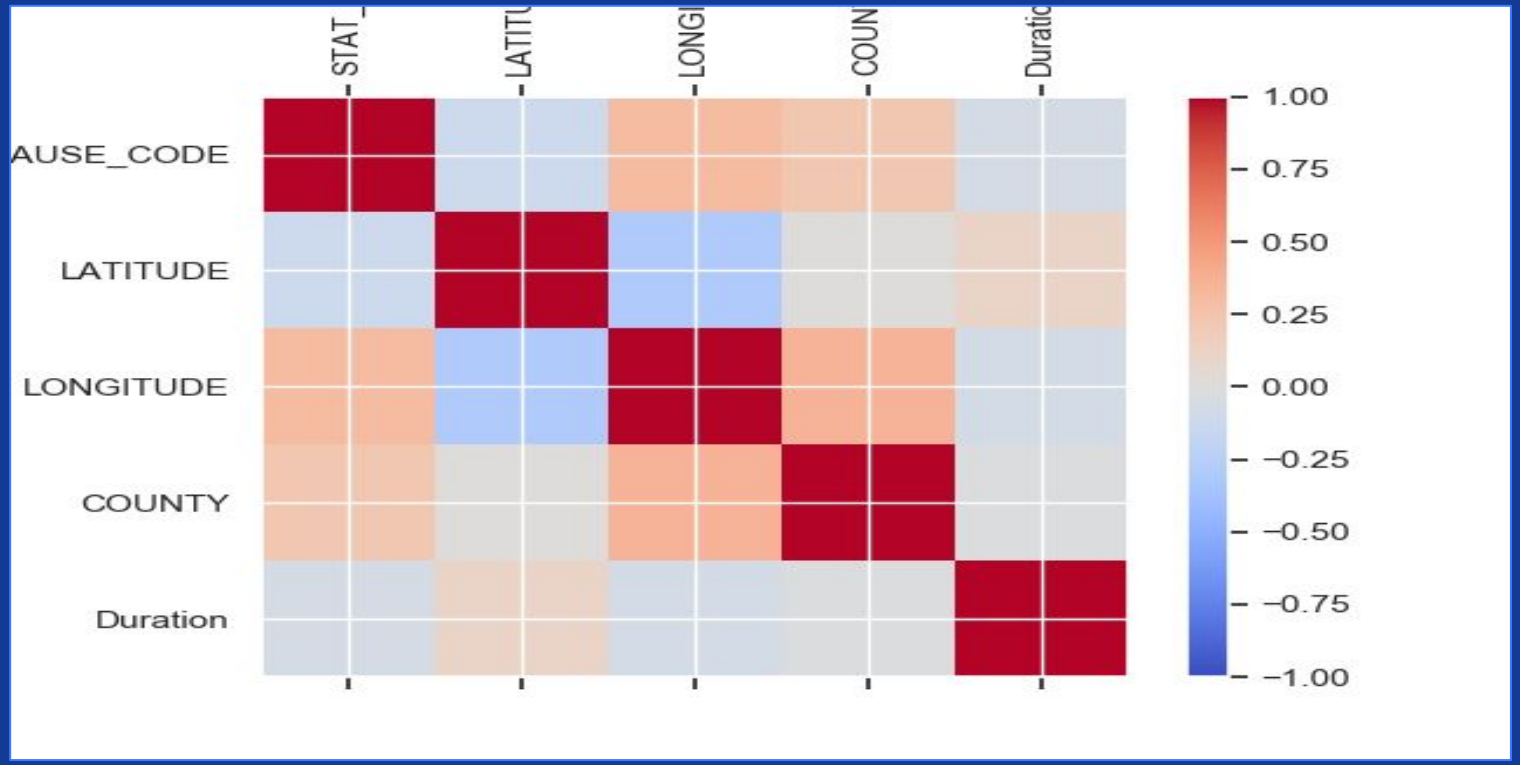
DATA PREPROCESSING

- Dropped Columns:"FIRE_YEAR","DISCOVERY_DATE","DISCOVERY_DOY","DISCOVERY_TIME","CONT_DATE",'CONT_TIME','STAT_CAUSE_DESCR','CONT_DAY','CONT_DOY'
- Created feature Duration using discovery day and containment day.

Classification Set



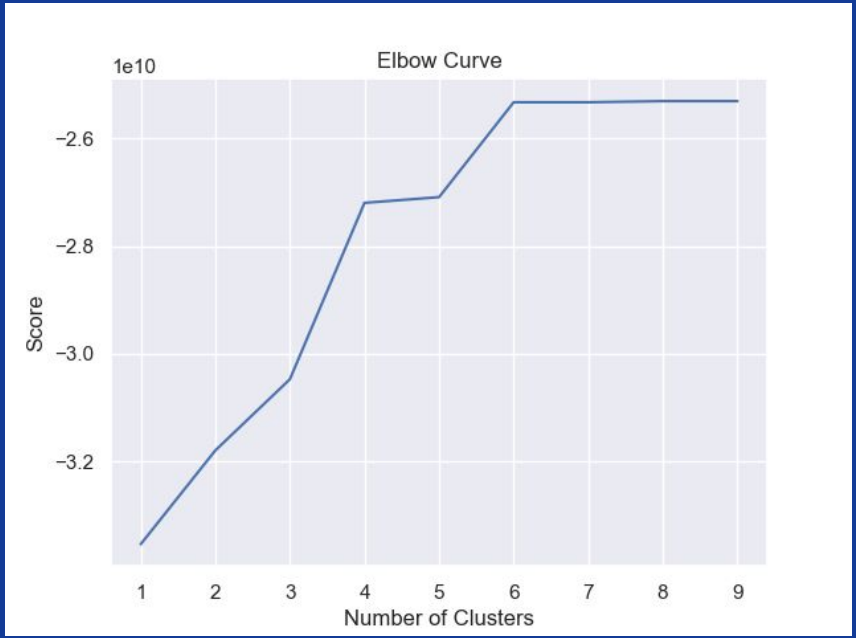
FEATURE SELECTION

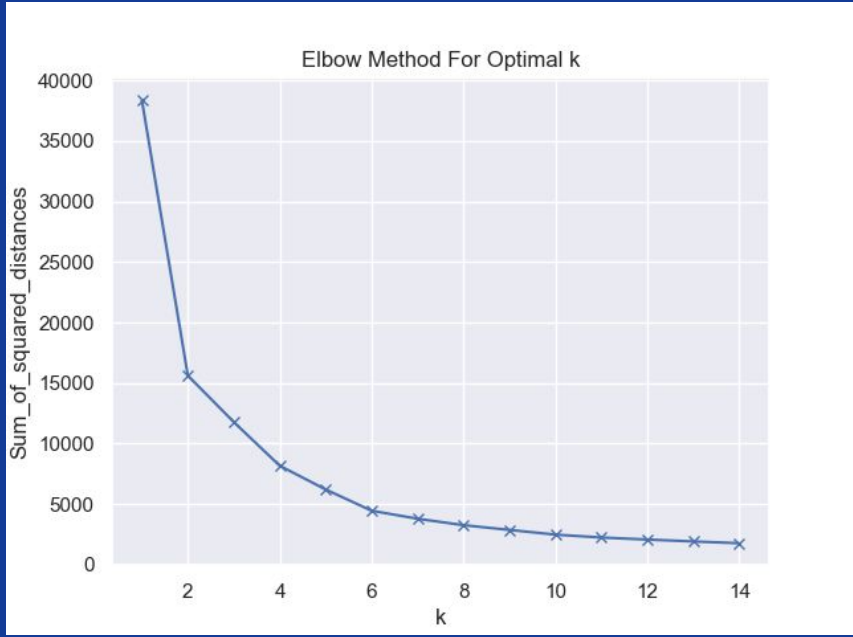


ALGORITHM 1: K-means

- The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (see below). This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.
- Number of clusters: 6
- Result:Clustering based on Lattitude, Longitude, Duration and Fire Size

Determining K



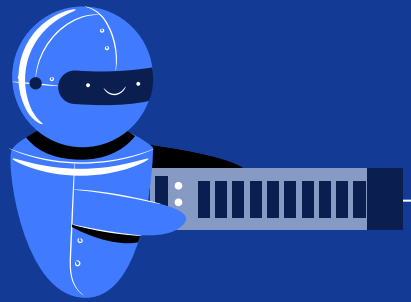


ALGORITHM 2: Random Forest

- Random forests train a set of decision trees separately, so the training can be done in parallel. The algorithm injects randomness into the training process so that each decision tree is a bit different. Combining the predictions from each tree reduces the variance of the predictions, improving the performance on test data.
- Configuration: 10
- Fire Size prediction accuracy: 78%
- Fire Cause prediction accuracy: 55%

03

RESULTS

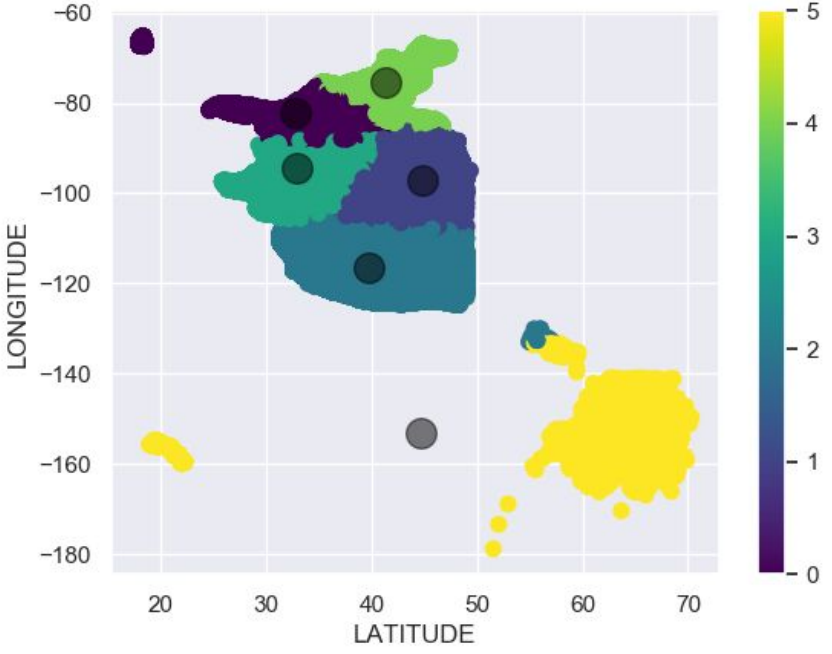


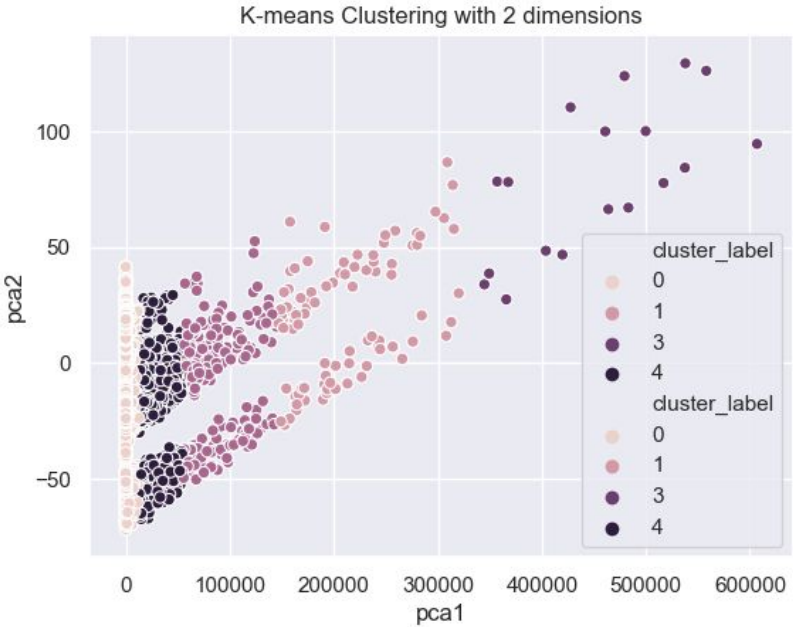
Classification

Objectives 1 & 2 were achieved using randomforests.

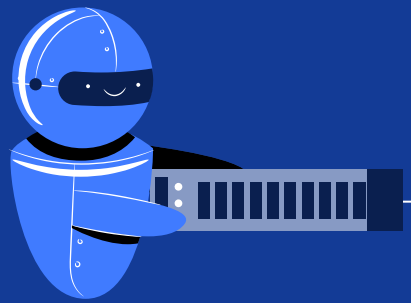
- Fire Size prediction accuracy: 78%
- Fire Cause prediction accuracy: 55%

Clustering


LONGITUDE
LATITUDE
−60
−80
−100
−120
−140
−160
−180
20
30
40
50
60
70
5
4
3
2
1
0


K-means Clustering with 2 dimensions
pca2
pca1
100
50
0
−50
0
100000
200000
300000
400000
500000
600000
cluster_label
0
1
3
4
cluster_label
0
1
3
4

04 CONCLUSIONS



CONCLUSIONS

- The proposed solution can be adopted for other natural disasters like earthquakes, hurricanes which depend on similar features etc.
- Limitations:
  - Not enough data
  - Unbalanced data
  - Empty values
- Future work:
  - Combining weather data and sensor data to provide real time analysis.

THANKS

Does anyone have any questions?