AN EXPLORATORY ANALYSIS AND PREDICTION OF FUEL TYPE

George Mason University

By Team 6

Abhishiek Kurra

Satyam Singh

Thanmayee Akkinen

Bharath Dindigala

Thanvi Malyala

AIT614
BIG DATA ESSENTIALS
SECTION 002

Advisor:
Dr. Duoduo Liao

INTRODUCTION

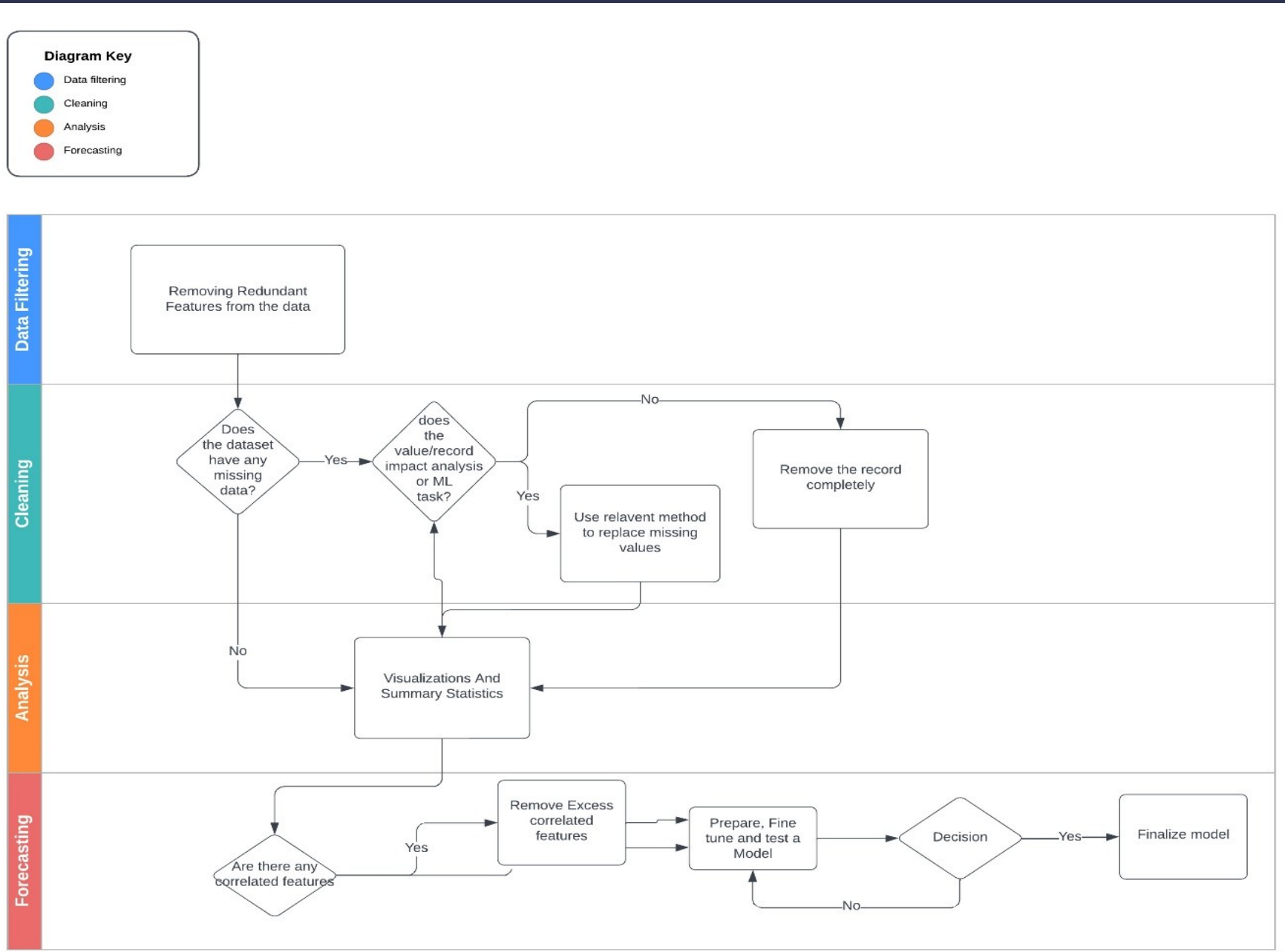
- It is now an established fact that Energy is the Epice Civilization (Smil, 2018). As such it is necessary to underst the amount of energy generated every yea
- **This is not a Project where we analyze and predict futu energy productions**. (This has already been done many tim over by many organizations
- For Analysis, we intend to analyze the energy g patterns across the years and among the types of the fuel type
- A Machine Learning Model for predicting the type of fu been made using the rest of the useful feature

THE DATASET

- The dataset is called “Global Powerplant Dataset” from The World Resource Institut
- Open-Source, Free to acces
- Just shy of 35000 records, each relating to a specific facilit
- 35 fields of data

country	country_	name	gppd_idi	capacity_	latitude	longitud	primary_	other_fue	other_fue	other_fue	commiss	owner	source	url	geolocati	wepp_id	year_of_	generatic	generatic	generatic	generatic	generatic	generatic	generatic	generatic	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimated_generation_note_
AFG	Afghanis	Kajaki Hy	GEODB(	33	32.322	65.119	Hydro						GEODB	http://glot	GEODB	1009793	2017									123.77	162.9	97.39	137.76	119.5	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
AFG	Afghanis	Kandahar	WKS007	10	31.67	65.795	Solar						Wiki-Sol	https://ww	Wiki-Solar											18.43	17.48	18.25	17.7	18.29	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
AFG	Afghanis	Kandahar	WKS007	10	31.623	65.792	Solar						Wiki-Sol	https://ww	Wiki-Solar											18.64	17.58	19.1	17.62	18.72	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
AFG	Afghanis	Mahipar	GEODB(	66	34.556	69.4787	Hydro						GEODB	http://glot	GEODB	1009795	2017									225.06	203.55	146.9	230.18	174.91	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
AFG	Afghanis	Naghlu D	GEODB(	100	34.641	69.717	Hydro						GEODB	http://glot	GEODB	1009797	2017									406.16	357.22	270.99	395.38	350.8	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
AFG	Afghanis	Nangarh	GEODB(	11.55	34.4847	70.3633	Hydro						GEODB	http://glot	GEODB	1009787	2017									58.77	54.42	42.71	59.72	46.12	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
AFG	Afghanis	Northwes	GEODB(	42	34.5638	69.1134	Gas						GEODB	http://glot	GEODB		2017														NO-EST	NO-EST	NO-EST	NO-EST	NO-EST-MATION	
AFG	Afghanis	Pul-e-Kh	GEODB(	6	35.9416	68.71	Hydro						GEODB	http://glot	GEODB		2017									21.99	21.19	18.4	25.34	19.74	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
AFG	Afghanis	Sarobi D	GEODB(	22	34.5865	69.7757	Hydro						GEODB	http://glot	GEODB	1009799	2017									123.23	82.87	69.15	93.83	80	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Bistrica 1	WRI1002	27	39.9116	20.1047	Hydro				1965		Energy D	http://www	GEODB	1021225										105.17	75.26	79.5	105.45	88.45	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Fierza	WRI1002	500	42.2514	20.0431	Hydro				1978		Energy D	http://www	GEODB	1021231										1976.01	1276.61	1503.72	1795.15	1648.24	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Koman	WRI1002	600	42.1033	19.8224	Hydro				1985		Energy D	http://www	GEODB	1021233										2072.13	1618.73	1805.63	2434.84	1982.72	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Lanabreg	WRI1002	5	41.3428	19.8964	Hydro				1951		Energy D	http://www	GEODB	1021236										20.37	12.89	14.64	20.04	15.23	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Shkopet	WRI1002	24	41.6796	19.8305	Hydro				1963		Energy D	http://www	GEODB	1021238										93.52	69.86	77.51	96.2	83.57	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Ulez	WRI1002	25	41.6796	19.8936	Hydro				1958		Energy D	http://www	GEODB	1021241										97.42	72.77	80.74	100.21	87.06	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Vau i Dij	WRI1002	250	42.0137	19.6359	Hydro				1971		Energy D	http://www	GEODB	1021242										895.02	561.94	614.47	897.47	703.64	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1	
ALB	Albania	Vlora	WRI1002	98	40.4874	19.434	Other						Energy D	http://www	GEODB	1021244															NO-EST	NO-EST	NO-EST	NO-EST	NO-ESTIMATION	
DZA	Algeria	Adrar	WKS006	20	27.908	-0.317	Solar						Wiki-Sol	https://ww	Wiki-Solar												35.22	34.22	35.33	35.17	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Ain Azel	WKS006	20	35.88	5.475	Solar						Wiki-Sol	https://ww	Wiki-Solar												38.68	37.56	38.37	38.75	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Ain Djas	WRI1023	520	35.8665	6.0262	Gas	Oil				SociÃƒÂ	Arab Uni	http://www	KTH	1069670														2171.28	NO-EST	NO-EST	NO-EST	NO-EST	CAPACITY-FACTOR-V1	
DZA	Algeria	Ain Sekh	WKS006	20	34.532	0.804	Solar						Wiki-Sol	https://ww	Wiki-Solar												34.85	33.67	34.54	35.46	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Ain el Ib	WKS006	20	34.346	3.164	Solar						Wiki-Sol	https://ww	Wiki-Solar												33.42	33.58	34.75	34.81	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Ain el Ib	WKS007	53	34.342	3.169	Solar						Wiki-Sol	https://ww	Wiki-Solar												80.98	81.84	85.66	85.55	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Ain el Ma	WKS006	20	34.861	4.204	Solar						Wiki-Sol	https://ww	Wiki-Solar												33.64	33.68	33.63	33.75	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	
DZA	Algeria	Algerie S	WKS006	43.5	27.908	-0.317	Solar						Wiki-Sol	https://ww	Wiki-Solar												73.79	72.11	74.36	74.02	NO-EST	SOLAR-'	SOLAR-'	SOLAR-	SOLAR-V1-NO-AGE	

ARCHITECTURE



PLATFORMS

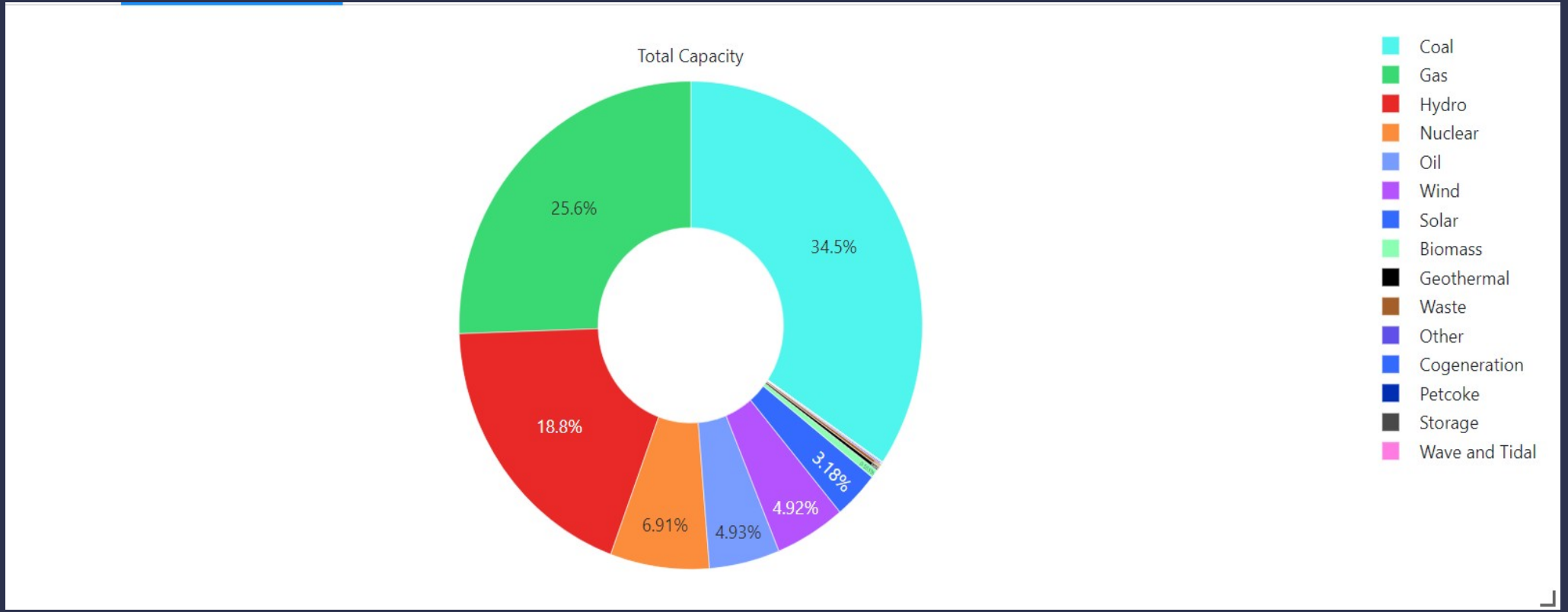
- We used the **Databricks** platform to perform the analysis and to m the ML Model
- Spark data frame with P package has been used to clean th data, train, test and visualize data.
- SQL has also been presenting data and making sc visualizations by using DataBricks inbuilt visualization tools

DATA PROCESSING AND METHODS USED

- The raw data imported from the csv file has 35000 rows of data related to power plant facilitie
- In these Redundant columns that take up additional space like url's source links and data citat have been removed
- The data pertaining to the Machine Learning methods have also been removed a unnecessary for a new ML model being built
- For Analysis, all the numeric columns are then converted into int and float respectively analysis was done on a random sample of 10000 rows of the data
- For Machine Learning, The data was further cleaned by removing all null values from the inte output column
- For the remaining numeric predictors, all the null values were set to
- For categorical predictors, the empty values were replaced with a string "Nul

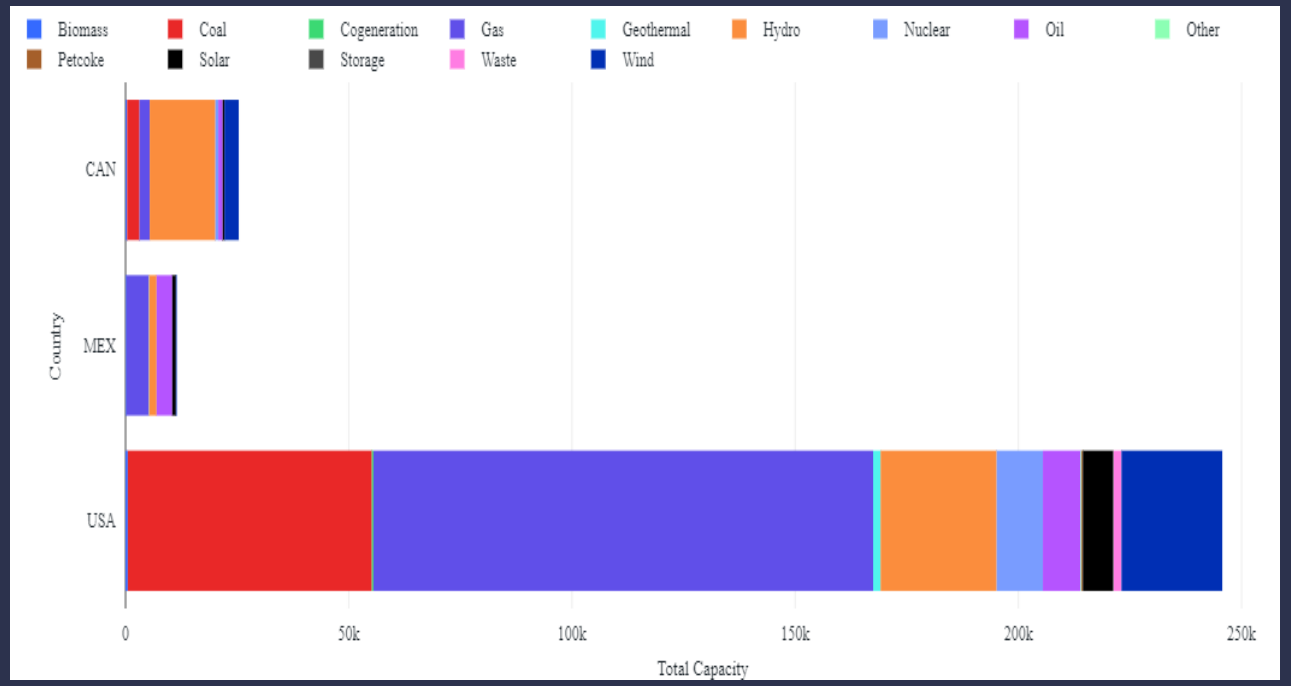
ANALYSIS RESULTS

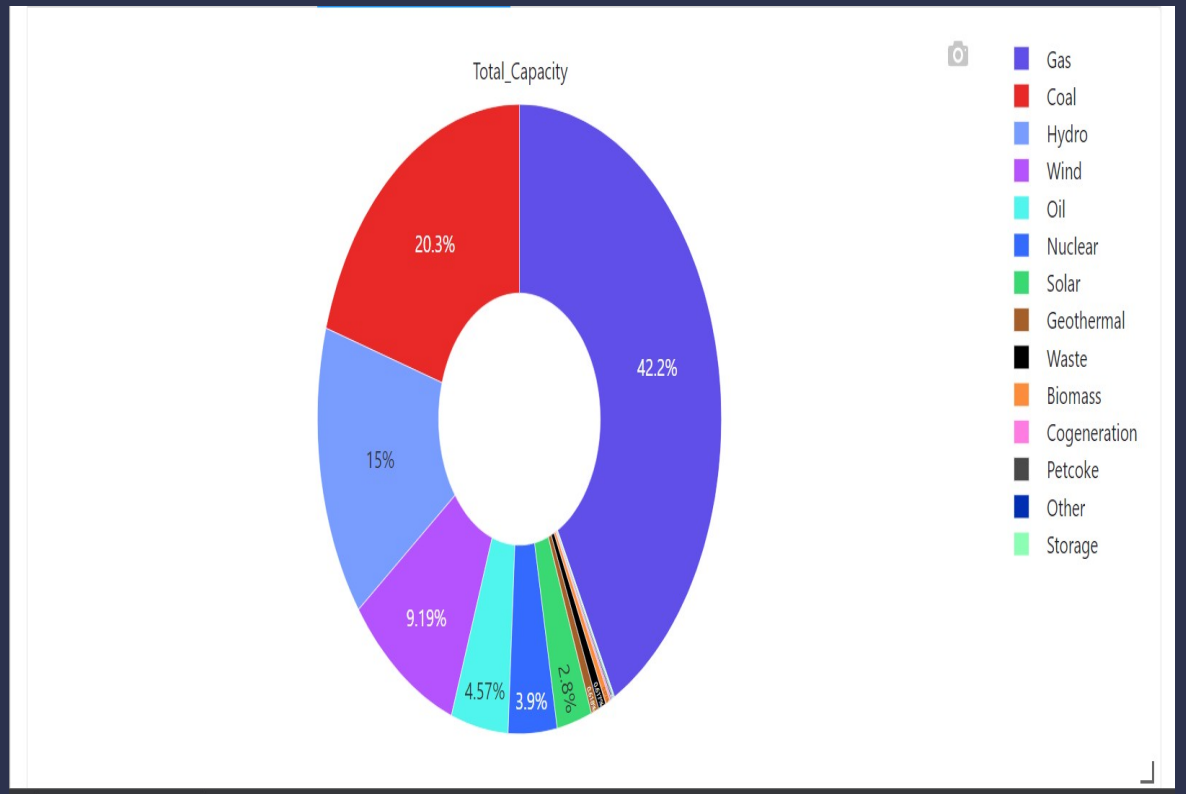
- what proportions of the world’s energy are eacl generated using each fuel type



ANALYSIS RESULTS

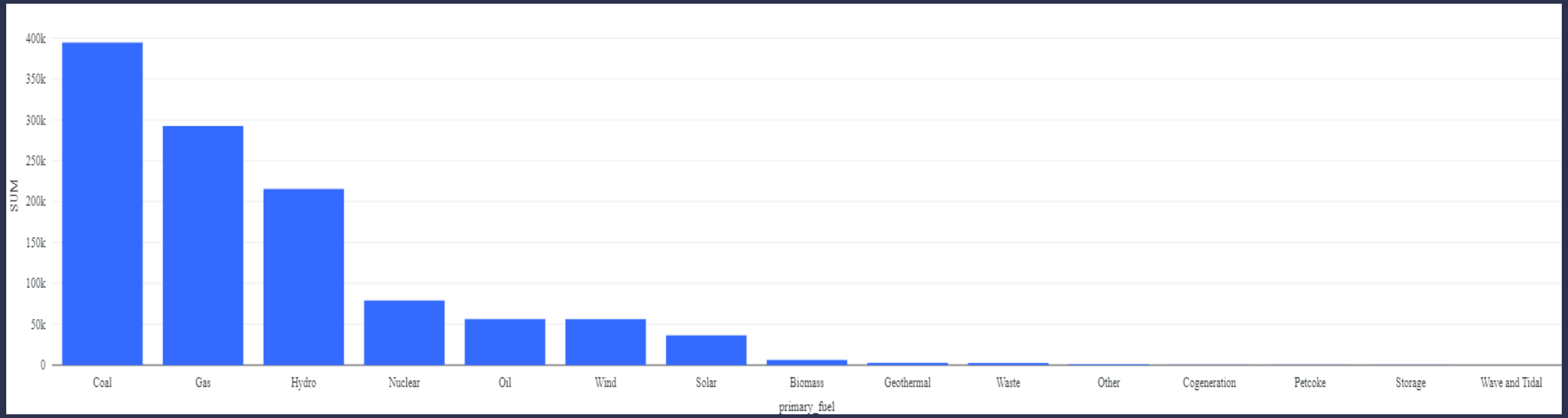
- What is the proportion of energy generation in North America?





ANALYSIS RESULTS

- Which type of plants tend to have higher plan capacities?



MACHINE LEARNING RESULTS

- Initial Fit’s Evaluation

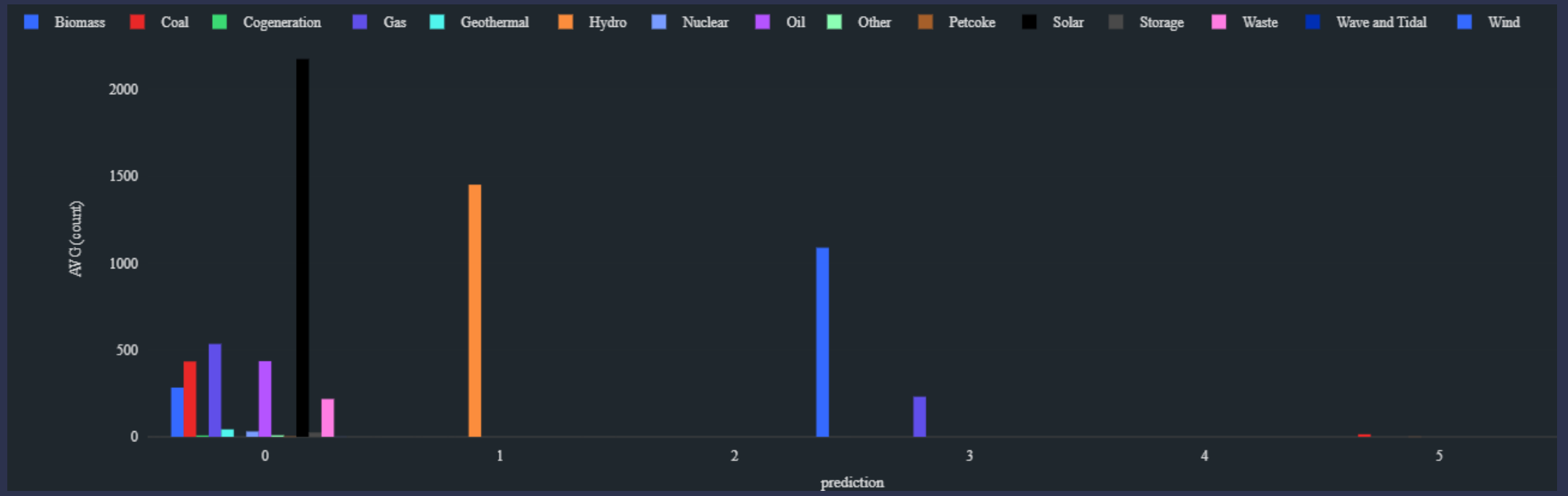
Area under ROC curve: 0.643605399792315

Accuracy: 0.839055793991416

- HyperTuned Fit’s Evaluation

Area under ROC curve: 0.999221207672384

Accuracy: 0.709871244635193



CONCLUSIONS

- The analysis of the dataset has given all of us an understanding of the situation c the power plants and energy generation during the years 2013 to 201
- Coal, Gas and Hydro energy generation plants amount to almost 79% of th global plant capacity
- A model has been prepared to predict the type of plant or primary fuel being use to generate the energy

FUTURE WORKS

- The dataset used here is pertaining to **70%** of the who **world's energy generation**, Finding the data for it and including it in the project would make f better model
- The use of a better model such as random forests could be suggeste
- Some of the features are simplified and stayed unused, these features can be u to analyse the integrity of the dat
- More features and predictors such as climate conditions, regional resources ca used to make better models
- Better hyper tuning parameters can be usec

ACKNOWLEDGEMENT

- We would like to thank Dr. Duoduo Liao (Professor) for providing feedbacks and guidance
- We thank Dr. Eddy Zhang (Professor) for their extensive guidance w project.
- We would also like to thank the Sai Deepak N. (Teaching Assistant) in provid us with guidance and informative resources
- We would also thank the George Mason University Faculty and Management providing us with the resources to successfully complete the projec

REFERENCES

- Smil, V. (2018) *Energy and civilization: A histor* . The MIT Press
- Jocelyn, V., & Biagi, I *Energy production in t* Statista. Retrieved March 16, 2 https://www.statista.com/study/48975/energy-production-in-the-united-states/?locale=en
- Jocelyn, V., & Biagi, *Global Electri* . Statista. Retrieved March 16 https://www.statista.com/study/74593/electricity-worldwide/
- *What is the Databricks File System (DB* What is the Databricks File System (DBFS)? | Databricks on AW Retrieved March 16, 2023, fron https://docs.databricks.com/dbfs/index.html
- *What is pyspark?: Domino data science dictio* . What is PySpark? | Domino Data Science Dictionary. (n.d.). Ret March 16, 2023, fron https://www.dominodatalab.com/data-science-dictionary/pyspark
- *What is Databri* What is Databricks? | Databricks on AWS. (n.d.). Retrieved Ma https://docs.databricks.com/introduction/index.html
- *Intelligent diagrammin* . Lucidchart. (n.d.). Retrieved March 16, 2023, fro https://www.lucidchart.com/pages
- *Simple gantt char* . Vertex42.com. (n.d.). Retrieved March 19, 2023, fro https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html?utm_source=ms&utm_medium=file&utm_campaign=office&utm_content=url