Cind 820: Big Data Analytics Literature Review

Adam Bruno
Student # 500904101
Dr. Tamer Abdou
Dr.Ceni Babaoglu
September 22, 2025

Abstract:

This literature review seeks to explore what relevant information can be obtained through the application of machine learning and classification models on pollutant levels from the UC Irvine Machine Learning Repository archives. Thorough review of past research has illustrated the significance of implementing simple systematic models, proper handling of data imbalances, as well as careful examination of evaluation metrics to validate and authenticate findings. This literature review will inquire on how past research and studies will shape the overall outlook of the project approach. Similarly, initial exploratory data analysis will be presented on the specified dataset demonstrating potential trends and relationships. Through EDA the project looks to visualize missing values data, as well as the existence of any potential anomalies to challenge the models. This literature review aims to provide a deeper understanding of the dataset as a baseline for further machine learning modelling. The review will investigate different classification models focusing on their strength and limitations. Each paper will be analyzed for its pertinence to the overall topic, legitimizing the integrity of the project through examination of similar experiments. Additionally, past research articles will be analysed to strengthen the project's position on the relevance of using classification techniques to evaluate and predict growing air pollutant levels. The literature review aims to clarify the project outlook by comparing it with relevant research articles.

Research questions:

1. Air pollutant quality can be measured through a wide range of contaminants, a large portion of this is mixed to create the conditions that numerous people inhabit everyday. As such, will the implementation of supervised learning models produce insightful discoveries into patterns and trends within modern living conditions?

2. How relevant and impactful are the identified discoveries from the model analysis? Is the data model sufficiently accurate can the model be used consistently or further scalable to address future issues? Which model has the highest potential to provide the most meaningful context without large computational necessities?

3. What distinctions can be made between the various models and approaches? Are there any overt advantages or limitations of using one specific model over another? How

does the EDA or data cleaning affect the models ability to perform? How much performance loss can be expected as a result of improper EDA implementation?

Literature review:

To better understand supervised learning and classifier models various researched articles were reviewed. These articles specifically covered utilization of modern algorithms to extrapolate and classify various ecological contaminants. The articles highlight various approaches to machine learning models as well as data processing. Review of past examples will assist in bolstering the overall integrity of the project. Similarly, the literature review attempts to explore the various distinct models in hopes of narrowing down an approach that best represents the dataset. These articles stress the importance of air quality index constraints that will utilise pollutant concentration variables such as PM2.5 to establish primary pollutant boundaries. Additionally, interpretation of the inherent relation between meteorological features and their impact on pollutant readings. This literature review will highlight past research experiences, illustrating how machine learning is becoming more prominent in the field of ecological studies.

Similar case studies can assist in approaching the question of whether machine learning could produce meaningful insights from air quality statistics. Liu et all (2024) highlights a distinct approach to analysing pollutant levels in the article *Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling.* The article delves into air pollutant variables present within Jinan, China from July 23 2020 through to July 13 2021. The research paper breaks down the fundamental steps to creating an effective predictive model, comparing various models to seek further proficiency. Firstly relevant data such as meteorological data and pollutant concentration data were obtained from the china air quality online repository during the previously specified time frame. The article illustrates the importance of proper data cleaning cautioning data validity loss due to external factors such as server outages. The article states “When building a prediction model, low-quality data may affect the prediction results. Therefore, it is necessary to clean the original data by eliminating noise and improve the data quality to improve the prediction accuracy.”(Liu et all 2024) After cleaning the data four models are tested including weighted model, light gradient boosted model, logistic regression model, and random forest model. The article further states the basic principles of each model specifically highlighting the LightGBM advantages including its high robusticity, accuracy, expandability and simple operation. It reports that the LightGBM demonstrated an accuracy rating of 97.5% as well as

f1 scores of 93.3%. Ultimately the article highlights the aptitude of utilising gradient boosted techniques in python to effectively forecast air quality.

Expanding on the topic of utilizing supervised machine learning models to extract evaluation metrics from rich variable dense sources, Liu et al (2015) exemplified similar research on the application of air quality forecasting within urban landscapes. The article, *Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rul*e, seeks to provide precise and reliable predictions to local authorities in Guangzhou. Similarly to the case study previously discussed, observational data was constructed within a dataset that included both meteorological parameters as well as pollutant levels. The article provides thorough insight into the use of neural networks to craft in depth training algorithms. Liu et al specifically states the approach's ability to model highly nonlinear functions and their ability to be trained for accurate generalizations. However it is also stated that three main factors can potentially limit a neural network training algorithm. These include network topology, learning algorithms, and learning samples. As such the paper aimed to develop a learning sample method that could effectively predict containment concentration levels based on a similarity principle of meteorology and pollutants. The paper combines the use of sensitivity experiments for parameter selection and Back Propagation neural network for dataset function computation. The model provided insightful results in forecasting performance of distinct pollutant elements. Evaluations within the slated May 2011 to April 2022 time frame demonstrated a 4% in MAPE values of the element PM10. This article ultimately shows the emergence of new forecasting algorithms to accurately depict pollutant data generalizations.

For the purpose of this literature review similar classification models will be analyzed to determine if additional approaches can provide insight into potential model limitations. The study conducted by Karthikeyan et al (2021) outlines the use of Naive Bayes and Support Vector machine learning models to predict potential climate crashes. The article specifically delves into oceanic climate simulations algorithms with regards to the parallel ocean program(POP2). Karithikeyan et al compares models to verify and quantify efficiency as a function of failure probabilities of the aforementioned POP2. Ultimately the paper aims to develop a deeper understanding of extreme weather event simulations. Additionally, the use of data extract techniques are validated as the research paper notes their improved results over standard meteorological approaches. The article provides insight into Naive Bayes algorithm theorem as the probability of B as P(A|B) as such calculating the probability of a

data point belonging to each class and assigning it to the class with the highest probability. Alternatively, the concept of Support Vector Machine models is also introduced as there are only two classes available: a binary classifier is used. This is used to ultimately identify a hyperplane within the divided data to predict straight forward. This is relevant to the literature review as it builds greater scope on the project's approaches by providing similar models performance. The results of the paper demonstrate that the Naive Bayes classifier produced an accuracy score of 0.9382 (93.82%). Ultimately the model correctly classified 151 of 156 failure cases. However, only classified 1 out of 5 success cases illustrating the models efficiency with the dominant failure cases. Alternatively, the SVM model yielded an accuracy score of 0.9691 (96.91%) , comparably reporting a score of 0.84280 when accounting for class imbalances. As such the paper depicts the SVM models superior performance highlighting its improved suitability for predicting extreme weather events. Overall the papers provide profound insight into the two distinct models highlighting machine learning continued emergence within the field of ecotoxicology. Additionally the model assists in singling out SVM as a clear model that this project could utilize for the dataset. This is due to the classifiers enhanced ability to handle imbalance or large datasets potentially offering greater limitation mitigation.

Data Source Description:

This project will be evaluating the dataset available from the UC Irvine Machine Learning Repository archives. The dataset contains records pertaining to pollutant levels within the specified station site and their years recorded. The project will mainly focus on Beijing and its abundant particle matter levels. Specifically, PM2.5 is atmospheric aerosol particles that leak into the air causing serious cardiovascular or respiratory issues if inhaled. These particles can originate from various emission sources such as vehicles, industrial facilities, wildfires, and natural events. The project examines historical data from January 1st 2010 to December 31st, 2015 to infer if any predictive insights are maintained. Similarly, missing values are present and denoted as NaN signifying data cleaning requirements. The dataset breakdown is as follows:

- No: row number
- Year: year of data recorded
- Month: month of data recorded
- Day: day of data recorded
- Hour: hour of data recorded

- Season: season of data recorded
- PM: PM2.5 concentration (ug/m^3)
- DEWP: Dew Point Temperature (ºC)
- TEMP: Temperature (ºC)
- HUMI: Humidity (%)
- PRES: Pressure (hPa)
- Cbwd: Combined wind direction
- Iws: Wind speed reading (m/s)
- Precipitation: hourly precipitation (mm)

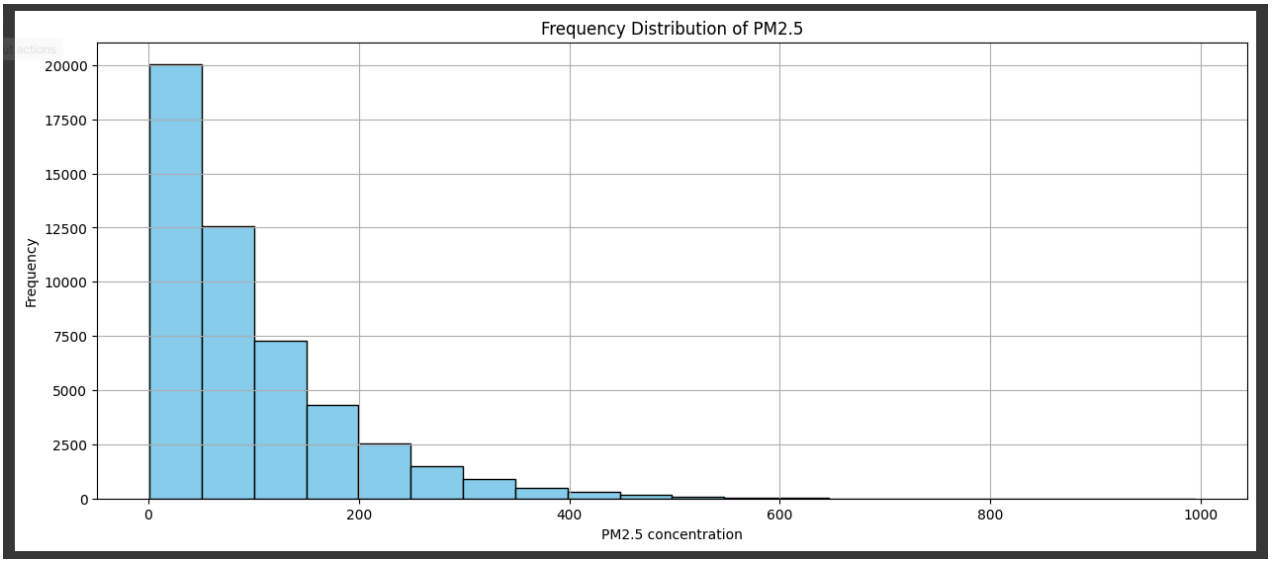
Documents will be downloaded as a csv file from the following repositories.

- PM2.5 Data of Five Chinese Cities - UCI Machine Learning Repository

EDA:

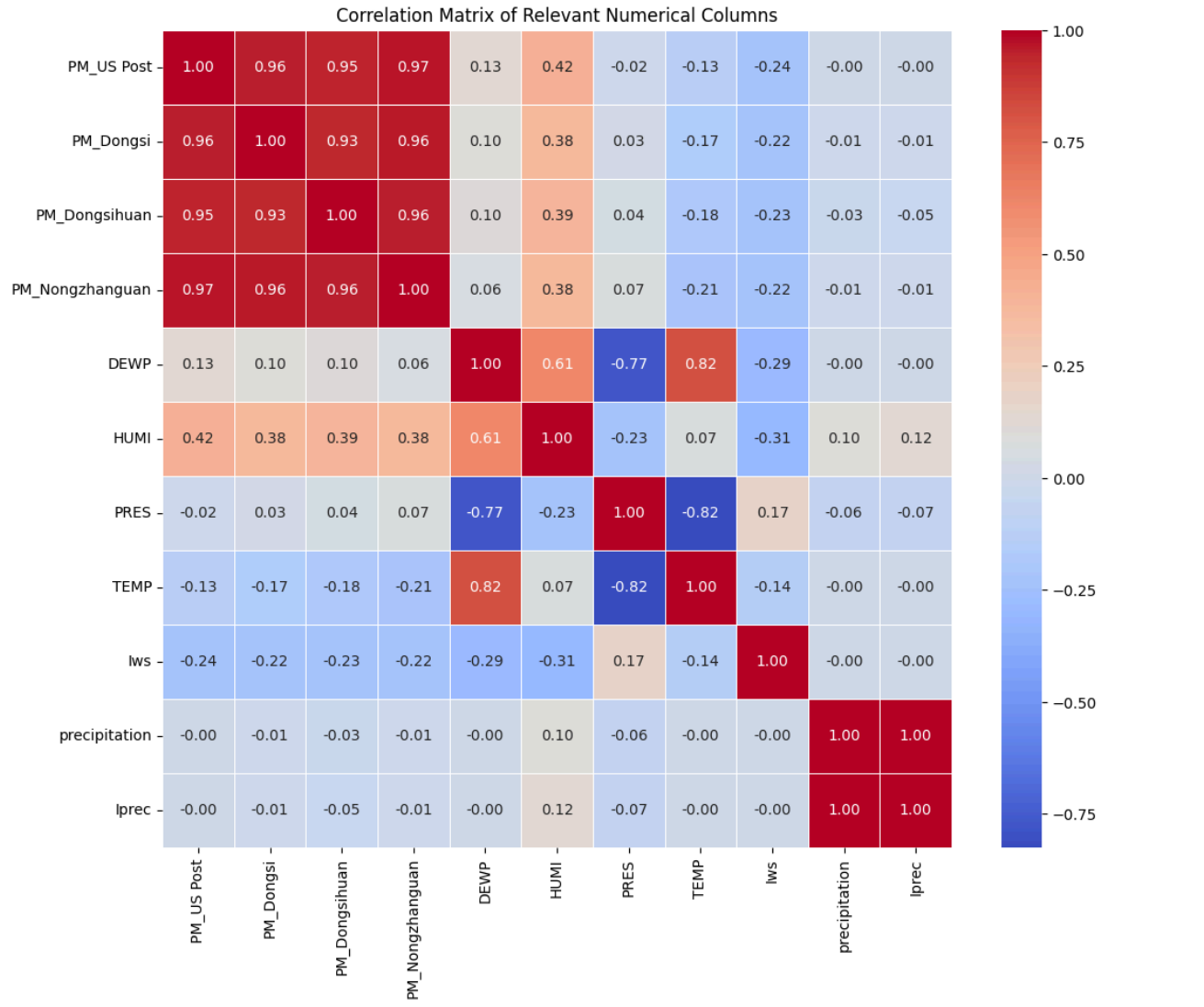
	No	year	month	day	hour	season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhanguan	PM_US Post	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Iprec
0	1	2010	1	1	0	4	NaN	NaN	NaN	NaN	-21.0	43.0	1021.0	-11.0	NW	1.79	0.0	0.0
1	2	2010	1	1	1	4	NaN	NaN	NaN	NaN	-21.0	47.0	1020.0	-12.0	NW	4.92	0.0	0.0
2	3	2010	1	1	2	4	NaN	NaN	NaN	NaN	-21.0	43.0	1019.0	-11.0	NW	6.71	0.0	0.0
3	4	2010	1	1	3	4	NaN	NaN	NaN	NaN	-21.0	55.0	1019.0	-14.0	NW	9.84	0.0	0.0
4	5	2010	1	1	4	4	NaN	NaN	NaN	NaN	-20.0	51.0	1018.0	-12.0	NW	12.97	0.0	0.0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52584 entries, 0 to 52583
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   No               52584 non-null  int64
 1   year             52584 non-null  int64
 2   month            52584 non-null  int64
 3   day              52584 non-null  int64
 4   hour             52584 non-null  int64
 5   season           52584 non-null  int64
 6   PM_Dongsi        25052 non-null  float64
 7   PM_Dongsihuan    20508 non-null  float64
 8   PM_Nongzhanguan  24931 non-null  float64
 9   PM_US Post       50387 non-null  float64
 10  DEWP             52579 non-null  float64
 11  HUMI             52245 non-null  float64
 12  PRES             52245 non-null  float64
 13  TEMP             52579 non-null  float64
 14  cbwd             52579 non-null  object
 15  Iws              52579 non-null  float64
 16  precipitation    52100 non-null  float64
 17  Iprec            52100 non-null  float64
dtypes: float64(11), int64(6), object(1)
memory usage: 7.2+ MB
None
```



No	0.000000
year	0.000000
month	0.000000
day	0.000000
hour	0.000000
season	0.000000
PM_Dongsi	52.358132
PM_Dongsihuan	60.999544
PM_Nongzhanguan	52.588240
PM_US Post	4.178077
DEWP	0.009509
HUMI	0.644683
PRES	0.644683
TEMP	0.009509
cbwd	0.009509
Iws	0.009509
precipitation	0.920432
Iprec	0.920432

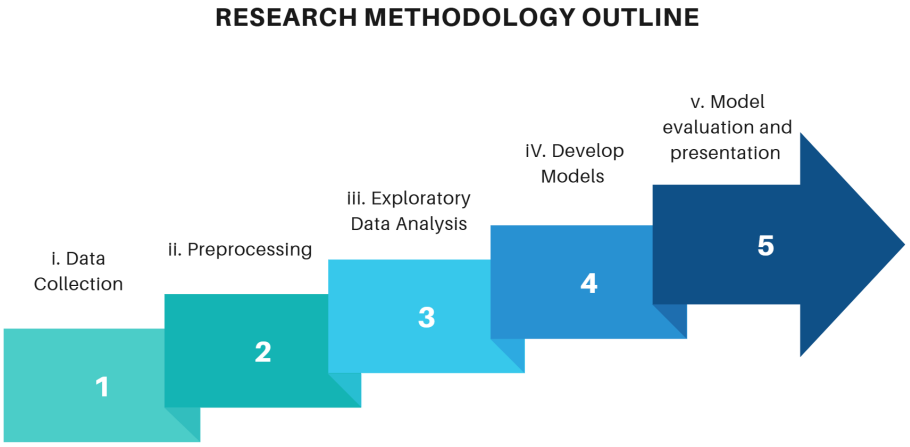
	No	year	month	day	hour	season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhanguan	PM_US Post	DEWP	HUMI	PRES	TEMP
count	52584.000000	52584.000000	52584.000000	52584.000000	52584.000000	52584.000000	25052.000000	20508.000000	24931.000000	50387.000000	52579.000000	52245.000000	52245.000000	52579.000000
mean	26292.500000	2012.499772	6.523962	15.726609	11.500000	2.491100	89.154439	92.560806	88.643737	95.904241	2.074554	54.602421	1016.465442	12.587040
std	15179.837614	1.707485	3.448452	8.798896	6.922252	1.116988	87.239267	88.027434	88.041166	91.643772	14.222059	25.991338	10.295070	12.098527
min	1.000000	2010.000000	1.000000	1.000000	0.000000	1.000000	3.000000	3.000000	3.000000	1.000000	-40.000000	2.000000	991.000000	-19.000000
25%	13146.750000	2011.000000	4.000000	8.000000	5.750000	1.000000	24.000000	28.000000	24.000000	27.000000	-10.000000	31.000000	1008.000000	2.000000
50%	26292.500000	2012.000000	7.000000	16.000000	11.500000	2.000000	64.000000	68.000000	62.000000	69.000000	2.000000	55.000000	1016.000000	14.000000
75%	39438.250000	2014.000000	10.000000	23.000000	17.250000	3.000000	124.000000	127.000000	122.000000	132.000000	15.000000	78.000000	1025.000000	23.000000
max	52584.000000	2015.000000	12.000000	31.000000	23.000000	4.000000	737.000000	672.000000	844.000000	994.000000	28.000000	100.000000	1046.000000	42.000000



Methodology:

This project will utilize various techniques and tools to establish an effective machine learning model on the specified dataset. The project will utilize python and its different libraries to establish and implement the machine learning model. The system of methods mentioned will be designed to provide scalability, performance, and interpretability. Data preparation and exploratory data analysis will involve libraries such as panda, numpy, and matplotlib to establish a general framework for the dataset. This serves to create a simple understanding of missing values as well as any data imbalances. Modelling libraries will

include sklearn and XGboost for developing the classifier models and machine learning algorithms. This project will implement gradient boosted techniques as aforementioned case studies have displayed its high performance capabilities. The model technique also boasts a high degree of scalability as well as improved accuracy scores. The methodology of this assignment will similarly follow a project outline that intends to offer deeper project scope. The graph illustrating the overall project methodology is as follows:



Github Repository Link:

[albruno-droid/CIND820: Data Analytics Air Quality Project]

References

Liu, Y., Zhu, Q., Yao, D., & Xu, W. (2015). Forecasting urban air quality via a back-propagation neural network and a selection sample rule. Atmosphere, 6(7), 891–907. https://doi.org/10.3390/atmos6070891

Liu, Q., Cui, B., & Liu, Z. (2024). Air Quality class prediction using machine learning methods based on monitoring data and secondary modeling. *Atmosphere*, *15*(5), 553. https://doi.org/10.3390/atmos15050553

Et. al., C. K. (2021). Prediction of climate change using SVM and naïve Bayes Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2). https://doi.org/10.17762/turcomat.v12i2.1856