PM2.5 concentration analysis and prediction in Chengdu

Srdjan Protic, protic783@gmail.com

I. Introduction

This project focuses specifically on Chengdu, China, utilizing data from the US Post location. The primary objectives include:

Linear Regression for PM2.5 Prediction in Chengdu: Developing a linear regression model to predict PM2.5 levels in Chengdu based on various environmental parameters. This analysis aims to understand how different factors contribute to air pollution in the specified location.

k-Nearest Neighbors (kNN) Classification for Chengdu: Implementing a kNN classification model to categorize PM2.5 levels in Chengdu into three pollution groups. This approach provides a classification framework to assess the severity of air pollution based on the similarity to neighboring samples.

By focusing on Chengdu at the US Post location, the project aims to provide insights into the specific air quality challenges in this area, aiding in the development of targeted strategies for pollution control and public health improvement.

II. Database

The database comprises data representing meteorological parameters in the city of Chengdu. We have a total of 52,584 samples, where each sample represents the measured values of meteorological parameters during one hour in a day.

There are a total of 17 different features available, including both categorical and numerical attributes. Categorical features relate to the measurement date, including hourly and seasonal data. There are 24 measurements per day.

The numerical features include common meteorological parameters such as temperature, humidity, air pressure, and wind direction. In addition to these, there are some unusual features such as cumulative wind speed (m/s), hourly precipitation (mm), and cumulative precipitation (mm). The most crucial numerical feature for our investigation is the concentration of PM2.5 particles (μg/m3) at the US Post location.

III. Exploratory Data Analysis

In the pursuit of refining the dataset for our scientific investigation, several thoughtful operations were executed to curate a dataset that is both pertinent and devoid of extraneous information. Each operation was undertaken with meticulous consideration, enhancing the overall quality and relevance of the data.

Exclusion of Features from Other Locations (Code snippet 1):

Two specific features, namely *'PM_Caotangsi'* and *'PM_Shahepu'*, originating from different locations, were judiciously omitted from our dataset. This discerning decision ensures that our analysis focuses exclusively on the distinctive characteristics of the *'PM_US Post'* location, aligning with the precise scope of our research.

Removal of Redundant Features - 'No', 'precipitation,' and 'Iprec' (Code snippet 1):

Further refinement was achieved by expunging features that, upon thoughtful examination, proved to be consistently negligible. The features 'No,' 'precipitation,' and 'Iprec' were deemed non-contributory due to their persistent proximity to zero.

In the pursuit of data integrity, a meticulous check for missing values was conducted. The results, presented as a percentage relative to the dataset size, serve as a testament to our commitment to transparency and data quality.

```
# Drop features from other locations
df.drop(['PM_Caotangsi', 'PM_Shahepu'], axis=1, inplace=True)

# Drop features 'precipitation' and 'Iprec'
df.drop(['No', 'precipitation', 'Iprec'], axis=1, inplace=True)

# Check for missing values
print(df.isnull().sum() / df.shape[0] * 100)
```

Code snippet 1: Drop irrelevant features and check for missing values

During the data analysis, a notable observation arises concerning the years 2010 and 2011, where over 85% of the data is missing for the 'PM_US Post' attribute. Given the critical significance of this particular feature to our investigation, samples with less than 15% annual data availability can be considered negligible. As a judicious measure, all data corresponding to these two years will be expunged.

Moving forward, the next crucial step entails the imputation of missing values for the remaining features. This process culminates in the creation of an enhanced and comprehensive database, poised for further investigation and analysis.

The histogram, presented in Figure 1, visually depicts the distribution of PM2.5 concentration in the dataset. By analyzing the histogram, we can gain insights into the distribution of PM2.5 concentrations, identify potential outliers, and understand the common concentration ranges within the dataset.

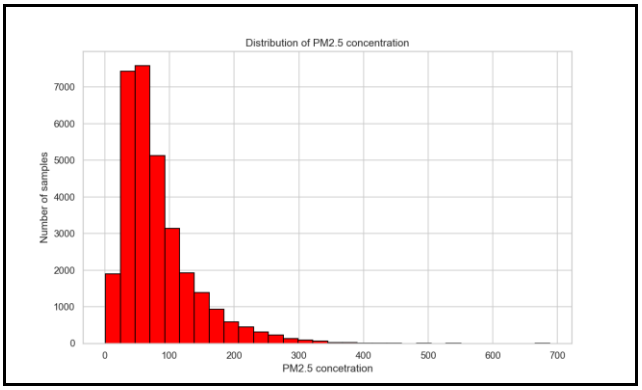

Figure 1: Frequency distribution of PM2.5 concentration in US_Post

The boxplot, presented in Figure 2, illustrates the distribution of PM2.5 concentration in the dataset, emphasizing the presence of outlier values. The horizontal box encompasses the interquartile range (IQR), with a line denoting the median concentration. Whiskers extend to 1.5 times the IQR from the quartiles, while outliers beyond this range are displayed as individual points. Notably, some values significantly deviate from the majority, particularly on January 31, 2014, coinciding with the Chinese New Year celebration. These outliers, representing exceptionally high particle concentrations due to festivities like fireworks, carry valuable information and should be retained in the dataset for a comprehensive analysis.

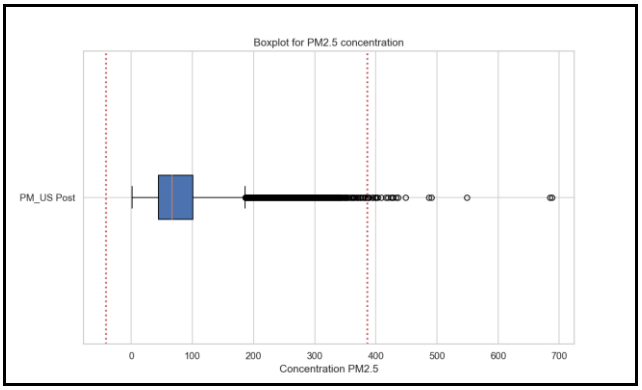

Figure 2: Frequency distribution of PM2.5 concentration in US_Post with focus on outliers

The diagram on Figure 3 represents the relationship between the seasonal variations and the concentration of PM2.5 particles in Chengdu is depicted. The x-axis denotes the four seasons (spring, summer, autumn, and winter), while the y-axis illustrates the concentration of PM2.5 particles in micrograms per cubic meter (μg/m³). Each point on the graph represents a sample from the database, where data was collected hourly throughout the day.

Upon analyzing the diagram, it becomes evident that the lowest concentration of PM2.5 particles is recorded during the summer months, depicted by the blue markers. This indicates that the air quality is typically better during the summer compared to other seasons. This information holds significance for understanding seasonal variations in air pollution and has potential implications for air quality management in Chengdu.rvi red svih paragrafa treba uvući za 3,6 mm. Za tekst u radu koristite Times New Roman 10 pt font.

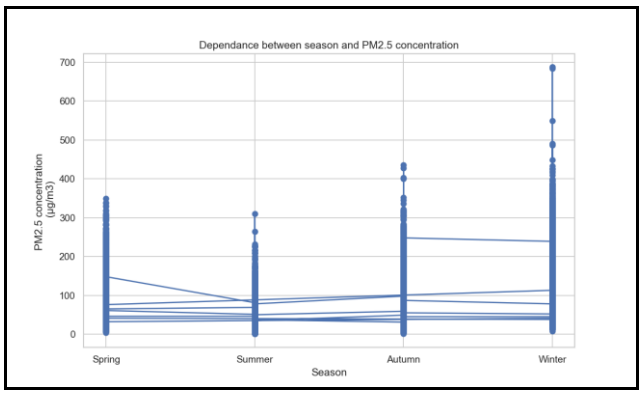

Figure 3: Distribution of PM2.5 Concentration Across Seasons

The diagram presented in Figure 4 illustrates the dependence between dew point (DEWP) and PM2.5 concentration in Chengdu. We can observe a certain pattern in the change of PM2.5 concentration concerning dew point values. This analysis contributes to understanding the influence of air humidity on air quality and adds to the overall assessment of factors affecting PM2.5 concentration in the investigated region.

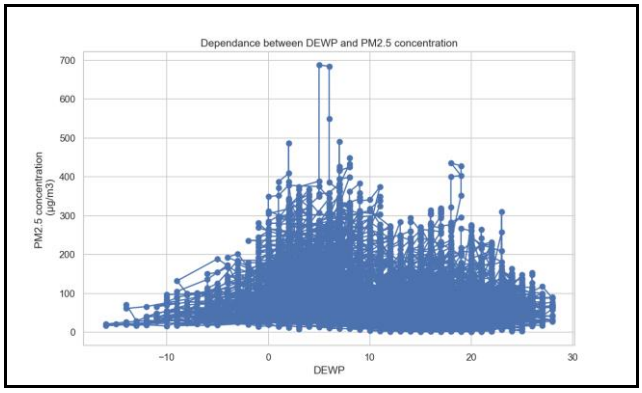

Figure 4: The Impact of Dew Point on PM2.5 Concentration

The dependencies between key meteorological factors and PM2.5 concentration in Chengdu are visualized in Figures 5, 6, and 7. Figure 5 illustrates the relationship between humidity (HUMI) and PM2.5 concentration,

highlighting the variability in PM2.5 levels with changing humidity. In Figure 6, the correlation between wind speed (Iws) and PM2.5 concentration is presented, revealing how variations in wind speed influence PM2.5 levels. Finally, Figure 7 explores the connection between temperature (TEMP) and PM2.5 concentration.

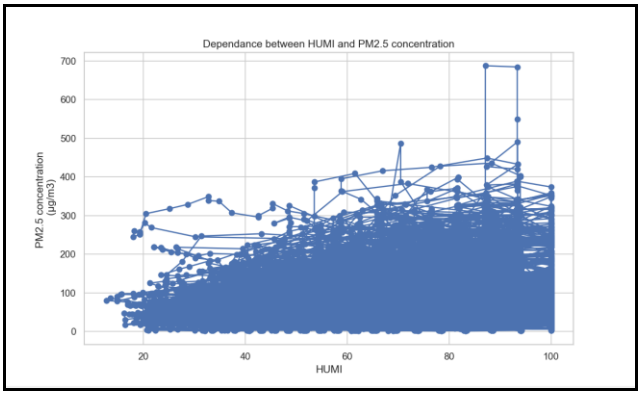

Figure 5: Dependance between HUMI and PM2.5 concentration

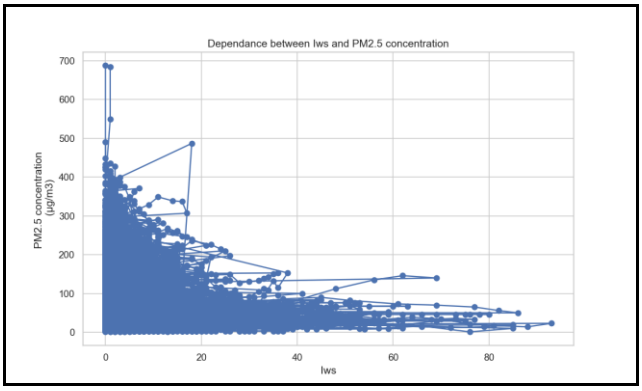

Figure 6: Dependance between Iws and PM2.5 concentration

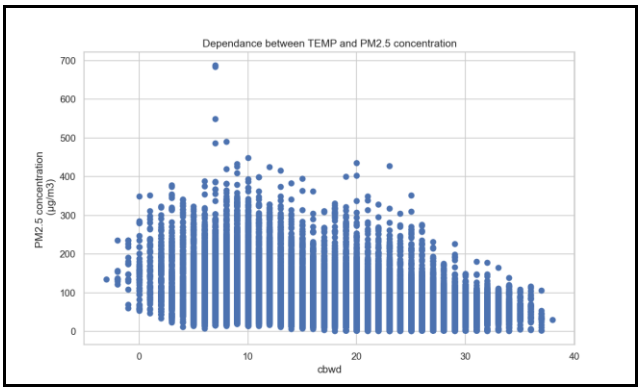

Figure 7: Dependance between TEMP and PM2.5 concentration

IV. Linear Regression

Linear regression models were tested using a methodology where 30% of randomly selected samples were allocated for the test set, while the remaining 70% of chosen samples were utilized for training the model.

This research delves into the efficacy of linear regression models in predicting the concentration of PM2.5 particles in the air. Initially, a linear regression model with a single hypothesis: *'y= b0+b1x1+...+bnxn'* was applied, but results showed insufficient accuracy (*Table 1*).

Model evaluation	Results
Mean squared error	2349.0535
Mean absolute error	35.2897
R2 score	0.2540
R2 adjusted score	0.2537

Table 1: Linear regression with hypothesis 'y= b0+b1x1+...+bnxn'

To enhance model performance, we introduced feature standardization, resulting with a no significant improvement.

We further explore the differences between two linear models using distinct hypotheses. While the second hypothesis:*'y=b0+b1x1+b2x2+...+bnxn+c1x1x2+c2x1x3 +...+d1x1^2+d2x2^2+...+dnxn^2',* yielded better results, it exhibited substantial variations in regression coefficients (*Table 2*).

This prompted consideration of Ridge regression to address overfitting and reduce coefficient disparities.

Model evaluation	Results
Mean squared error	1677.1612
Mean absolute error	30.0104
R2 score	0.4675
R2 adjusted score	0.4653

Table 2: Linear regression with hypothesis 'y=b0+b1x1+b2x2+...+bnxn+c1x1x2+c2x1x3+...+d1x 1^2+d2x2^2+...+dnxn^2'

This research provides insights into the application of linear regression models in predicting PM2.5 particle concentrations. Feature standardization and the adoption of Ridge regression proved pivotal in improving model performance and reducing variations in regression coefficients. Results are represented on *Figure 9* and *Table 3*.

Model evaluation	Results
Mean squared error	1676.6744
Mean absolute error	30.0007
R2 score	0.4675
R2 adjusted score	0.4653

Table 3: Ridge regression

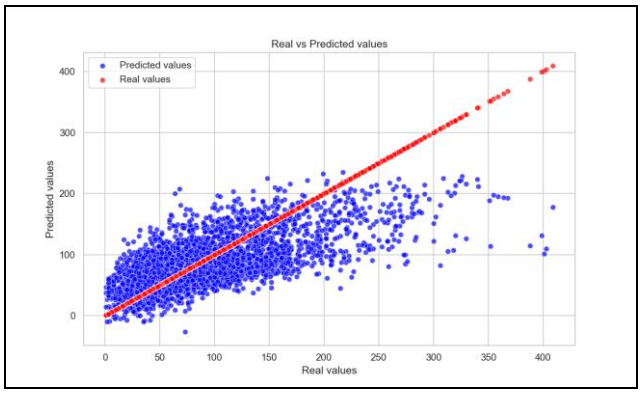

Figure 9: Ridge regression

V. KNN CLASSIFIER

Classification of hazard levels using the k-Nearest Neighbors (kNN) algorithm involves making decisions about the classification of a specific PM2.5 feature into one of the classes based on the class membership of its nearest neighbors in the training set. In the context of this methodology, the kNN algorithm assesses the similarity between the feature in question and its surrounding data points by measuring distances in the feature space. The decision on the class membership is then determined by the majority class among the k-nearest neighbors.

This approach leverages the idea that similar instances in the feature space tend to belong to the same class. By employing the kNN algorithm, we aim to capture local patterns and dependencies within the data, allowing for a flexible and adaptive classification of PM2.5 hazard levels.

In the categorization of PM2.5 levels into distinct hazard categories, three classifications are established based on the 'PM_US Post' values. The 'Safe' category encompasses instances where the 'PM_US Post' value is less than 55.5, resulting in the assignment of **'safe'** to the 'danger' column. Moving into the 'Unsafe' category, 'PM_US Post' values falling between 55.4 and 150.4 (inclusive) prompt the assignment of **'unsafe'** to the 'danger' column. It is important to note that the upper limit for 'unsafe' is set at 150.4, and the lower limit is 55.5. Finally, the **'dangerous'** category involves 'PM_US Post' values surpassing 150.4, leading to the assignment of **'dangerous'** to the 'danger' column. This systematic approach provides a clear classification of PM2.5 levels into safety zones, facilitating an understanding of potential health risks associated with varying concentration levels.

In our case, we will use cross-validation with 10 subsets. By applying cross-validation for each combination of parameter values k, weights, and metric, the optimal performance of the kNN classifier in terms of classification accuracy is achieved with *Manhattan* metric, k set to *12*, and weights parameter set to *distance*.

Based on the final confusion matrix obtained by accumulating matrices from each of the 10 cross-validation iterations, performance metrics for the classifier are computed, including accuracy, precision, recall, specificity, and F-measure.

Model evaluation was conducted on an independent test set, providing insights into the actual algorithm performance.

Based on the presented results in *Table 4*, a comprehensive conclusion can be drawn, affirming the robust performance of our PM2.5 hazard classification model. The amalgamation of precision, accuracy, sensitivity, specificity, and F-score metrics for each defined class (Safe, Unsafe, Dangerous) showcases exceptional model capabilities. Particularly noteworthy is the model's outstanding precision in classifying 'Safe' and 'Unsafe' categories, with precision, sensitivity, and F-score values approaching unity. The remarkable specificity in classifying the 'Dangerous' category underscores the model's reliability in identifying genuinely hazardous situations. The overall accuracy of the model stands out as exceptionally high, affirming its efficacy in PM2.5 hazard classification. These results contribute to solidifying the reliability and effectiveness of our model, providing a sturdy foundation for further research and real-world applications.

	Safe	Unsafe	Dangerous
Precision	0.9867	0.9867	0.9834
Accuracy	0.9914	0.9863	0.9948
Sensitivity	0.9867	0.9867	0.9834
Specificity	0.9917	0.9860	0.9980
F score	0.9867	0.9867	0.9834

Table 4:kNN Classifier