Swinburne University of Technology
COS30049 Computing Technology Innovation Project
Semester 2, 2024

Tutorial Class: Tuesday 8:30am to 10:30am
Tutor: Qian Li

Assignment 2

Exploration and Implementation of Machine Learning Models

Anti-Pesto Party

Henry Richardson	104 420 453
Seth Kalantzis	103 992 935
Matthew Cross	101 828 627

Contents

Introduction

This report and accompanying study aims to explore the relationship between pollutants and respiratory health outcomes in the state of New South Wales. By analyzing data captured by the NSW Air Quality Monitoring Network alongside reports from Local Health Districts, machine learning models can reveal patterns and provide valuable insights. The findings from this study aim to help NSW residents and policy makers understand how air pollution may contribute to/affect pre-existing respiratory health conditions, and inform their decision making on their respiratory well-being.

Problem Framing

The main objective of this study is to analyse the impact of air pollutants on respiratory health to help residents and policy makers make informed decisions. While it is generally understood that poor air quality is linked to respiratory conditions, the relationships between specific pollutants and health conditions aren't easily definable. Existing solutions may rely on simple statistical methods, but a machine learning approach is more appropriate to find patterns in complex datasets. The specific goals for the study are:

1. Establish a baseline understanding/correlation between specific air pollutants and health outcomes. Quantify the relationship between air quality and health risks. (Linear Regression)
2. Identify which specific combination of pollutants are most relevant to health outcomes, and assess their impact on the relevant health statistics. (K-Means)
3. Assess and predict risk levels based on air quality and particle count. Identify situations where respiratory health issues are mostly likely to spike. (K-Nearest Neighbors)

Data Collection

Data collection serves as the foundation for subsequent decision-making and analysis in the machine learning lifecycle. It is imperative that collected data contains validated features with predictive power in relation to the study's objective. This study evaluated the quality of raw data based on the reputability of the source, the consistency and reliability of the collection instruments, and how representative and abundant the data was.

Air quality features were collected from the NSW Air Quality Monitoring Network and provide insight into the quantities of six pollutants within the air at monthly intervals from 49 Air Quality Collection Sites. Health outcome labels were extracted from nine separate xls files provided by Health Stats NSW. Data was collected from hospitals within NSWs 15 Local Health Districts (LHDs) and covers all available outcomes relevant to respiratory health.

The mapping of Air Quality Collection Sites to LHDs proved difficult as no information on suburbs within the LHDs was available. Map data was manually entered in Google Maps and exported to a csv file to provide mappings for the processing phase. Additionally, due to the size of the air quality dataset, the 24-years of data was exported in 4-year segments to avoid gateway timeout with their website. It was later recombined during preprocessing.

Data Processing

The data for the models was prepared by merging eleven datasets from three different sources. The following steps were taken to ensure consistency and relevance.

Process Individual Datasets

Exploratory analysis helped to inform initial preprocessing activities. Column names were normalized and irrelevant data like '*confidence interval*' and '*wkt*' (geographic location) were dropped to reduce noise. Bayesian Ridge Regression was used to fill missing values as it provided reliable predictions for the cyclical pollutant level data. Outliers were set to the appropriate upper and lower bounds, and duplicate rows and columns were removed from the dataset. In the Air Quality dataset specifically, 229 columns representing different combinations of suburb and pollutant were combined into six pollutant columns and multiple rows.

Merge Datasets

Air quality and health datasets were integrated using an inner joining to retain all relevant data. Missing pollutant data for certain health districts was imputed using linear regression. The merged datasets were then sorted by '*financial year*' and '*lhd*' as per their common columns. The data was then aggregated annually and monthly, and split into gendered and genderless subsets.

Final Processing for Different Models

Final preprocessing was tailored to each model. For Linear Regression, the data was normalized and split 80/20. For K-Means, non-essential columns were dropped, and pollutants and health statistics were scaled. For KNN, a '*pollution score*' was derived by summing pollutant values and scaling them, then weighted to calculate a '*health risk score*' before splitting for training and testing 80/20.

Machine Learning Model Selection

Models were selected based on the specific questions being answered and the characteristics of the collected datasets, including the relatively small size, the data types, and the diversity of features and labels. K-Means Clustering, Linear Regression and K-Nearest Neighbor (KNN) were selected as the primary models for the project. ARIMA Time-Series Forecasting was also explored to enhance the project's ability to forecast health and air-quality outcomes over time. This supplementary analysis is outlined in Appendix A.

K-Means Clustering

Clustering is an unsupervised ML technique which groups unlabelled data based on their similarities, and was identified as the most appropriate approach to identifying which specific combination of pollutants are most relevant to health outcomes. A K-Means clustering model with Principal Component Analysis (PCA) was chosen to address this problem due to its computational efficiency, and the ease of interpretation for relatively small datasets. By optimizing features through PCA, selecting the optimal number of clusters via the Elbow Method, and uncovering hidden patterns with the K-means model, clusters could be identified and analysed based on a silhouette analysis, calinski-harabasz scores, and davies-bouldin scores.

Alternatives considered included DBSCAN and Hierarchical Clustering. DBSCAN is appropriate for identifying non-linear clusters and handling noise points, but requires careful tuning of parameters which proved challenging for the smaller datasets. Hierarchical Clustering proved to be less scalable and provided fewer insights when plotted to a dendrogram than K-Means did when plotted to a 3D Scatter Plot.

Linear Regression

Regression techniques model the relationship between dependent and independent variables to predict continuous outcomes. This project used Linear Regression to assess the relationship between specific air pollutants and health outcomes given its simplicity, fast training time, and the assumed linear relationship of the dataset. Preliminary analysis utilized Scatterplots to check feature correlations and Residual Plots to confirm homoscedasticity, ensuring consistent variance in errors. The model's transparency made it ideal for exploring multiple hypotheses and avoiding overfitting.

Linear Regression was selected over alternative approaches such as Polynomial Regression due to the assumed linear relationship of the data and the clarity of analysing outcomes. Tree-based regression models such as Decision Trees and Random Forest were not selected as they have a higher risk of overfitting with smaller datasets.

K-Nearest Neighbor Classification

KNN is a non-parametric algorithm that classifies data by finding the '*k*' nearest neighbors. This method is effective for smaller datasets since it doesn't require extensive training and makes predictions based on proximity to existing data points. KNN Classification was utilized to determine risk levels for geographic regions based on pollution levels. Preliminary analysis informed model optimization by experimenting with various values of '*k*', and using cross-validation to identify the best fit. The model's accuracy, precision, recall, and classification results were assessed to ensure robust predictions over other models.

An alternative considered was The XGBoost Classifier. However, the model proved more complex in its implementation than KNN without significant improvements in predictive power or interpretability of results.

Technical Implementation

The selected ML models were implemented with Python, utilizing libraries such as Scikit-learn, Pandas, Numpy, Matplotlib, Statsmodels, IPython and KNeed. Scikitlearn provided the necessary tools for clustering and supervised learning, while Statsmodel was used for ARIMA time-series forecasting. Data preprocessing, including cleaning and normalization, was done using Pandas and Numpy.

Reference materials were sourced from library documentation, as well as supplementary sources such as Kaggle, GeeksforGeeks, and DataCamp. These materials provided code templates and tutorials for setting up and training the models.

Key challenges included pivoting away from initial research completed on housing market prices due to lack of data found in the data collection phase. Refocusing on air quality and health outcomes provided more meaningful data, but necessitated substantial data transformation and manipulation during the preprocessing phase. Ensuring consistency and model compatibility by feature scaling proved challenging, but resulted in accurate and precise model performance. Finally, optimizing hyperparameters required significant experimentation, especially in the clustering and classification models.

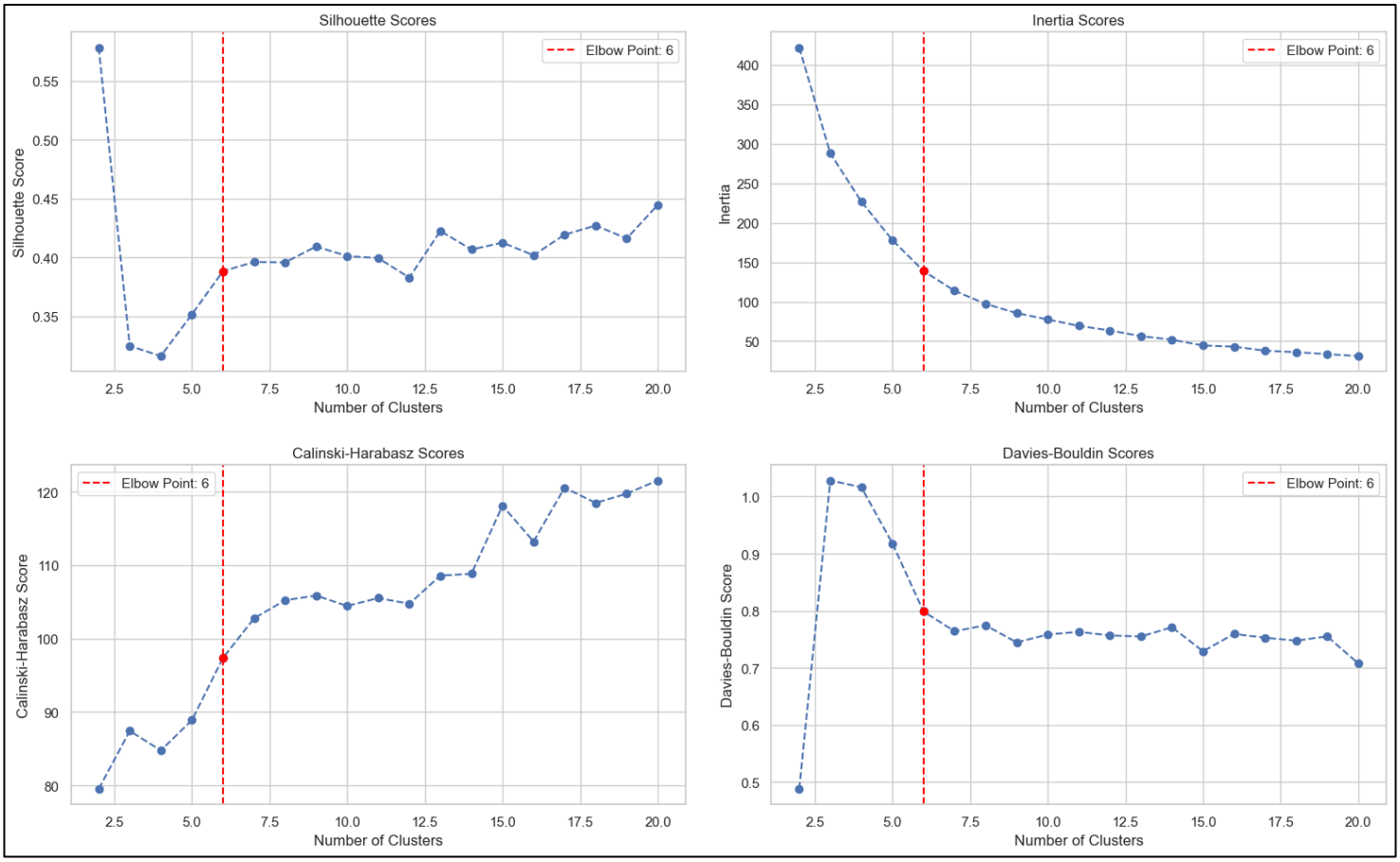
The final implementation differed from the initial expectations due to the necessity of PCA dimensionality reduction to enhance the K-Means model, and the time-consuming cross validation required to fine-tune models to avoid over and underfitting risks. Finding a good balance in hyperparameters and data usage to avoid under or overfitting due to our dataset's size and high feature dimensionality also proved difficult.

Implementation Evaluation

K-Means Clustering

K-Means clustering uncovered the strongest results with the annual data. Initial Principal Analysis (PCA) found that 88% of the variance in the pollutants could be explained by three components. The optimal number of clusters was identified as six using the elbow method. This level of clustering produced a Silhouette Score of 0.39, Inertia of 138.95, a Calinski-Harabasz Score of 97.37, and a Davies-Bouldin Score of 0.80 (Figure-1). The moderate Silhouette Score is offset by the relatively strong Davies-Bouldin and Calinski-Harabasz Scores, indicating decent compactness and separation across the clusters.

Figure-1. Evaluation metrics for varying cluster sizes in K-Means analysis.



A 3D visualisation of components (Figure-2) and a Silhouette Analysis (Figure-3) indicate that the trained model is able to separate Clusters 3-5 with higher accuracy. The strength of this analysis is highlighted by the Asthma Emergency Department Hospitalisation rates per 100,000 of females (Figure-4) and males (Figure-5) when mapped to the clusters. The boxplots, histograms and metrics tables indicate much lower distribution of results.

Figure-2. 3D Scatter Plot of Clusters by PCA Components.

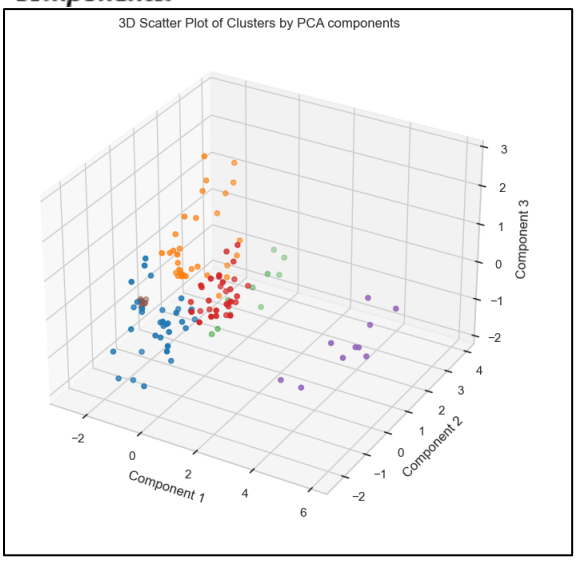


Figure-3. Silhouette Analysis of Clusters.

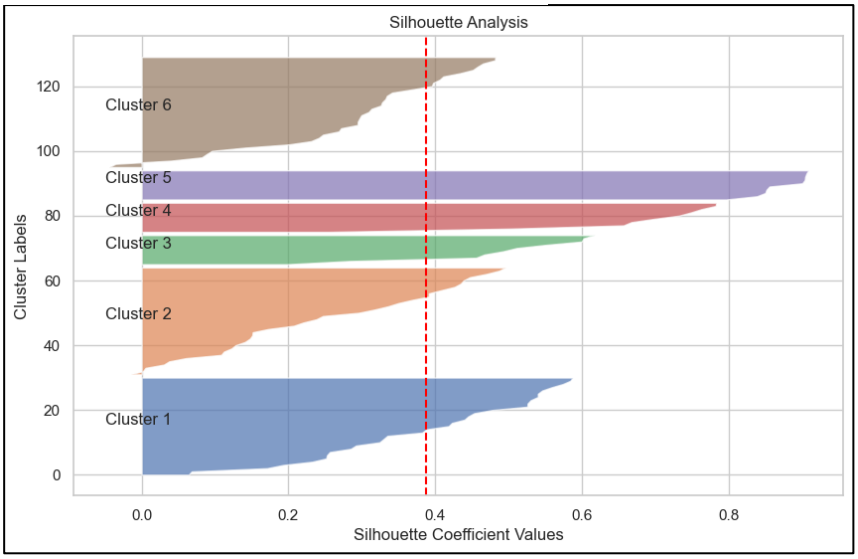


Figure-4. Asthma EDP Female mapped to Clusters: Boxplot and Histogram.

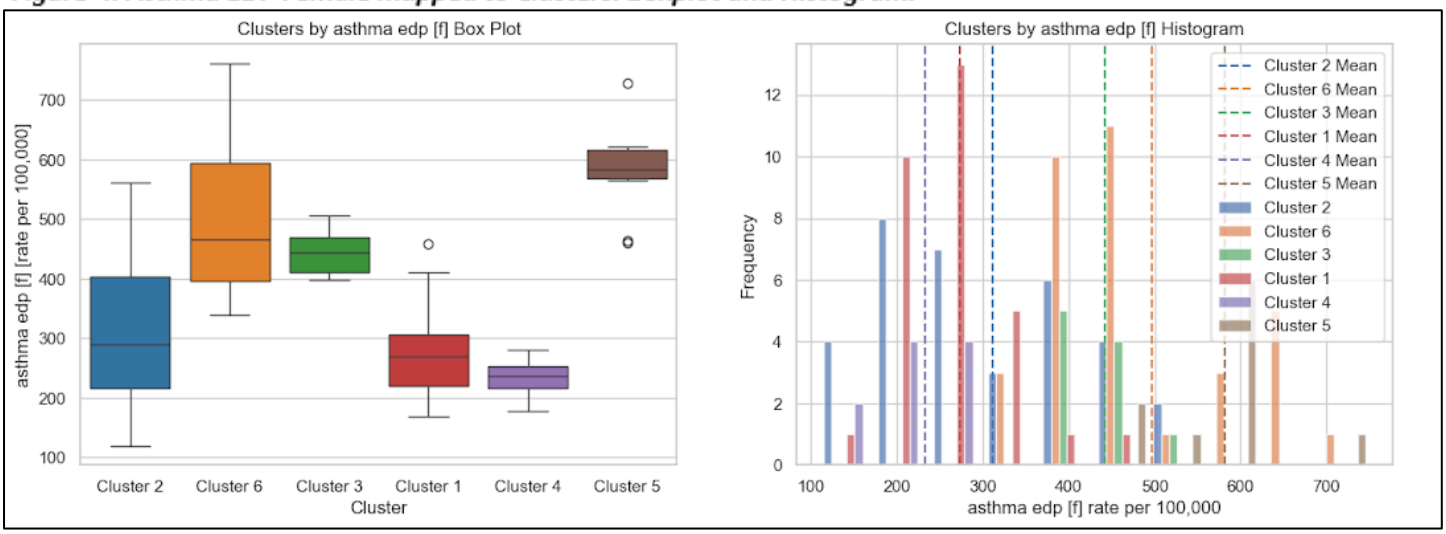
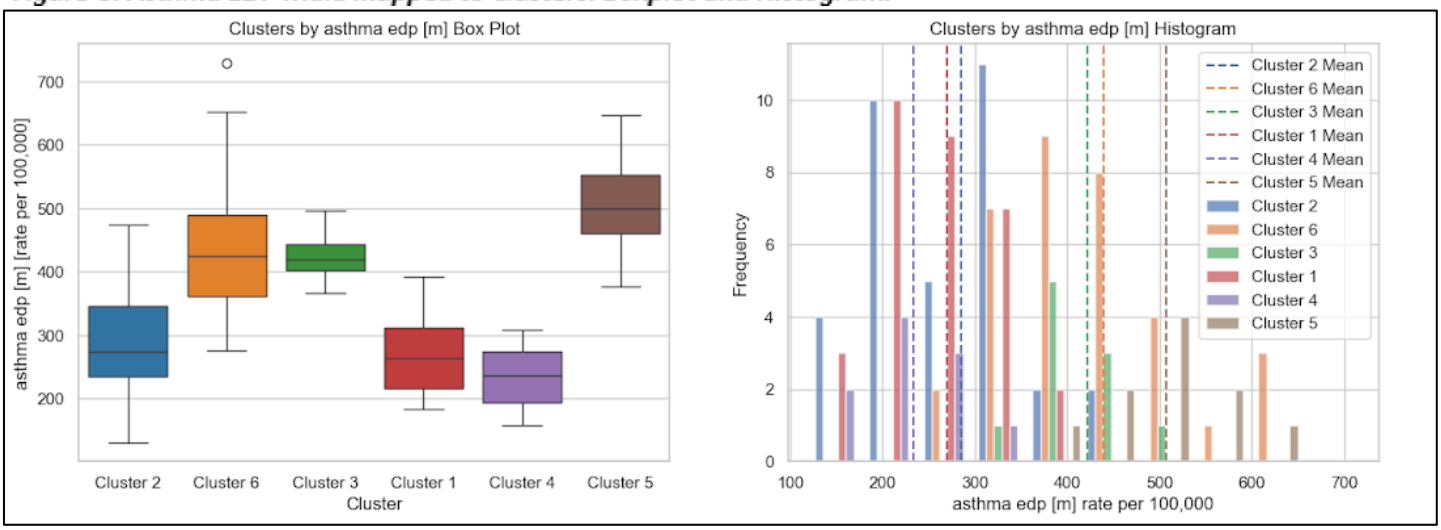


Figure-5. Asthma EDP Male mapped to Clusters: Boxplot and Histogram.

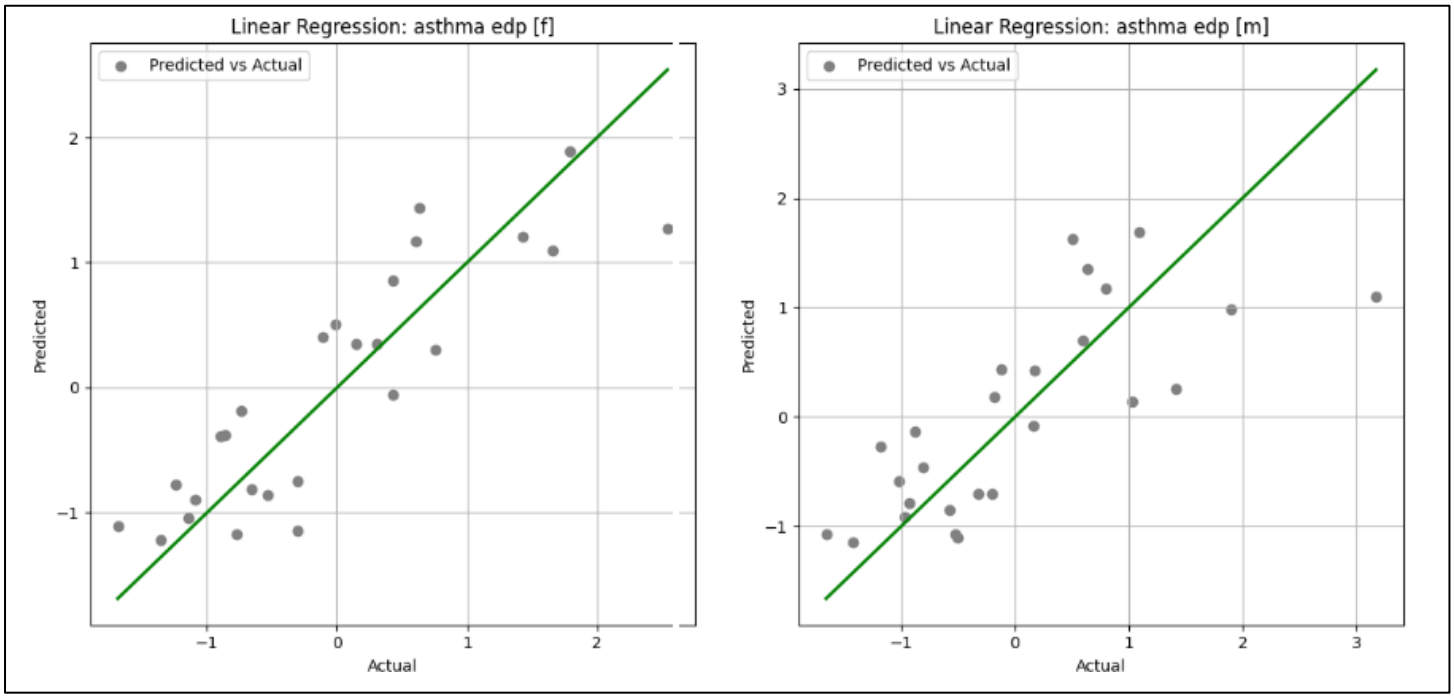


Linear Regression

The Linear Regression Model generally found low linear correlation. The model's performance was validated using Mean Squared Error, Mean Absolute Error, and several other metrics. The model was able to uncover a relationship between Asthma-related emergency presentations (r^2 = 0.76 in women, 0.57 in men) (Figure-7), and suggested that OZONE was the leading factor in health outcomes among the pollutants. Though these findings don't address the overall problem of assessing health outcomes, they do provide useful context for further analysis.

Other regression models were subsequently tested, but each provided no further insights. Ridge

Figure-6. Asthma EDP Male and Female Linear Regression Analysis.



Regression presented near identical results, while Random Forest and Polynomial Regression suffered from overfitting.

K-Nearest Neighbor Classification

The aim of KNN was to predict the health risk levels (low, medium, high) based on pollutant data and Asthma EDPs. In an example test case, the model achieved an impressive accuracy of 96.8% Figure-7), indicating the model performed exceptionally in correctly classifying the risk levels. '*Medium Risk*' and '*High Risk*' showed strong precision and recall; however, '*Low Risk*' was slightly lower at 0.86, indicating occasional misclassification (expected due to lower data points of '*Low Risk*' class).

Figure-7. KNN Analysis Metrics.

Accuracy: 0.967948717948718

	precision	recall	f1-score	support
Low Risk	1.00	0.86	0.92	21
Medium Risk	0.96	0.97	0.97	148
High Risk	0.97	0.98	0.98	143
accuracy			0.97	312
macro avg	0.98	0.94	0.96	312
weighted avg	0.97	0.97	0.97	312

Cross validation was used to determine the optimal k-value of 13, providing the best performance. Data scaling ensured features were comparable, which is critical for KNN. The confusion matrix (Figure-9) showed few misclassifications, mostly occurring between '*Low*' and '*Medium*', with the Decision Boundary plot demonstrating the same (Figure-8).

Figure-8. KNN Decision Boundary with Data Points.

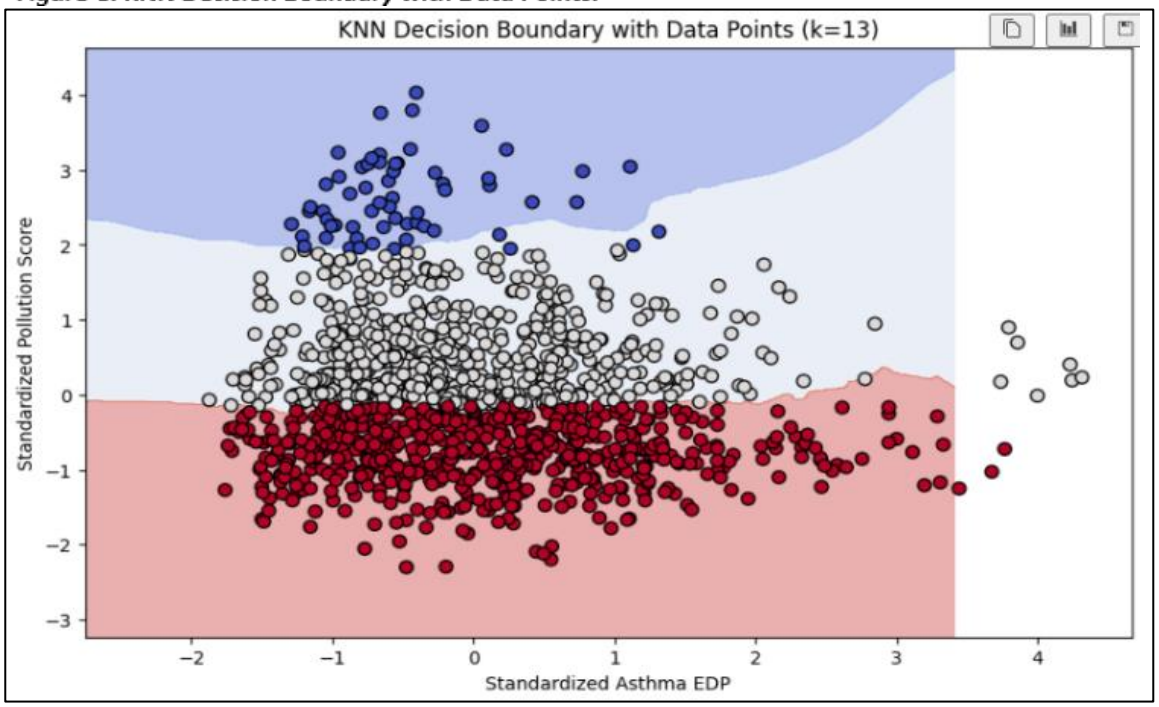
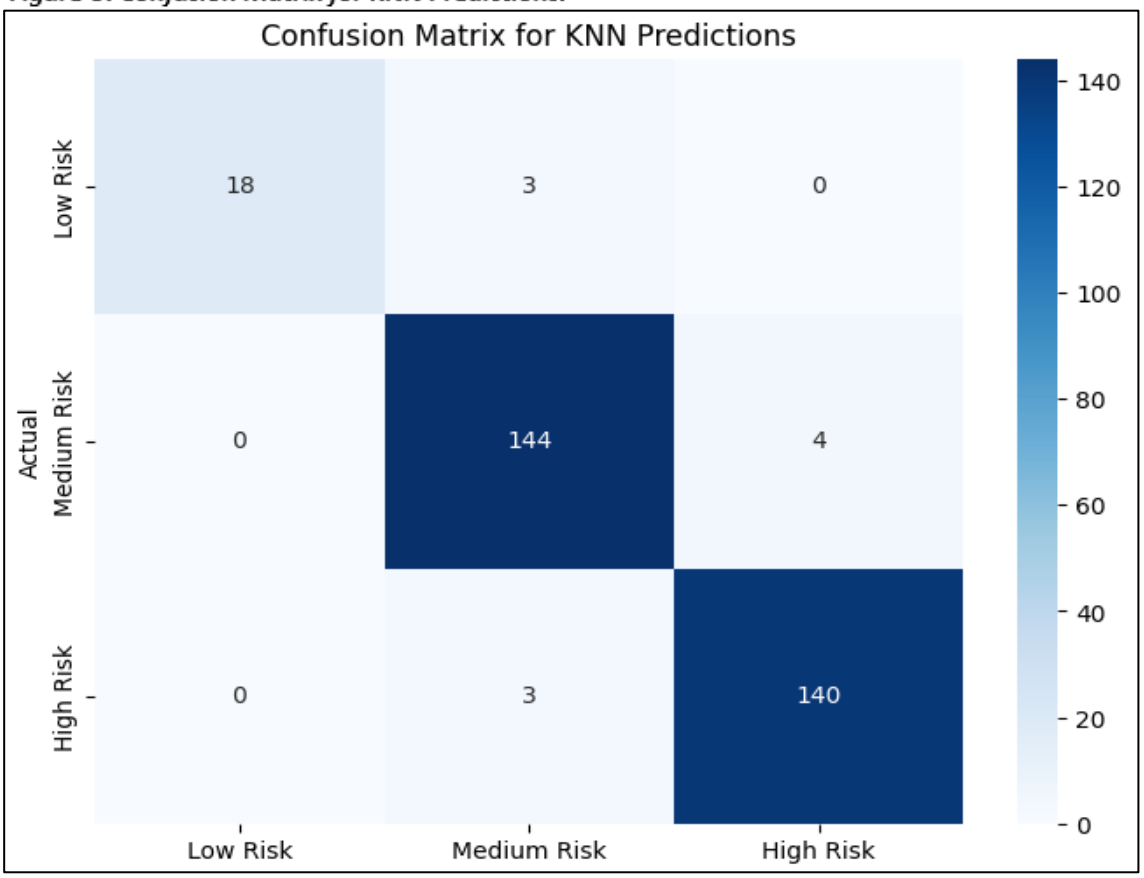


Figure-9. Confusion Matrix for KNN Predictions.



A major unexpected outcome was how well the model performed, achieving a high accuracy and precision. Given the nature of the problem and the small dataset, it was expected that the model would struggle with the less frequent '*Low Risk*', however it still performed well even there.

Overall, the model successfully addresses the problem of predicting health risk based on pollution data, demonstrating high accuracy and balanced performance across risk levels.

Conclusion

This report explored the relationship between common air pollutants and respiratory health outcomes for residents of NSW by training various machine learning models. The K-Means clustering model identified groupings of pollutants most impacted health outcomes. Linear regression analysed each individual pollutant and health outcome to find correlations and make predictions. KNN classification then effectively classified health risk levels.

The results in this report have important implications in informing the decision making of NSW residents to mitigate personal risks, as well as health authorities when setting public policy. The results also indicate that machine learning models can reveal patterns and provide valuable insights where traditional techniques are limited.

The moderate correlations identified by the report indicate that other features beyond the six air pollutants measured are required to deepen the analysis. Future work would benefit from the inclusion of additional datasets related to factors such as socio-economic trends in order to account for the variance in health outcomes not explained by air quality.

This report serves as a guide for identifying which air pollutants most significantly impact respiratory health outcomes.

Bibliography

Agrawal, R 2021, *Evaluation Metrics for Your Regression Model,* Analytics Vidhya, viewed September 23 2024, <https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/>.

Altexsoft 2023, *Data Collection for Machine Learning: Steps, Methods, and Best Practices*, Altexsoft, viewed 19 September 2024, <https://www.altexsoft.com/blog/data-collection-machine-learning/>.

Aronson, L 2019, *ARIMA Modeling and Train/Test Split*, GitHub, viewed 26th September 2024, <https://laurenliz22.github.io/arima_modeling_and_train_test_split>.

Aronson, L 2019, *ARIMA Modeling and Train/Test Split*, GitHub, viewed 26th September 2024, <https://laurenliz22.github.io/arima_modeling_and_train_test_split>.

Brownlee, J 2023, *How to Create an ARIMA Model for Time Series Forecasting in Python, Machine Learning Mastery*, viewed 26th September 2024, <https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>.

Brownlee, J 2023, *How to Create an ARIMA Model for Time Series Forecasting in Python*, Machine Learning Mastery, viewed 26th September 2024, <https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>.

Dabbura, I 2018, *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*, Medium, viewed 25 September 2024, <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.

Depth, A. R. 2019, *KNN Visualization in just 13 lines of code*, Medium, viewed 27th September 2024, <https://towardsdatascience.com/knn-visualization-in-just-13-lines-of-code-32820d72c6b6>.

Frost, J 2017, *Overfitting Regression Models: Problems, Detection, and Avoidance - Statistics By Jim,* Statistics By Jim, viewed September 24 2024, <https://statisticsbyjim.com/regression/overfitting-regression-models/>

Frost, J 2018, *How To Interpret R-Squared in Regression Analysis,* Statistics By Jim, viewed September 23 2024, <https://statisticsbyjim.com/regression/interpret-r-squared-regression/>.

GeeksforGeeks 2023, *Davies-Bouldin Index*, GeeksforGeeks, viewed 27 September 2024, <https://www.geeksforgeeks.org/davies-bouldin-index/>.

GeeksforGeeks 2023, *Regression Metrics,* GeeksforGeeks, viewed September 22 2024, <https://www.geeksforgeeks.org/regression-metrics/>.

GeeksforGeeks 2024, *ARMA TIME SERIES MODEL*, GeeksforGeeks, viewed 26th September 2024, <https://www.geeksforgeeks.org/arma-time-series-model/>.

GeeksforGeeks 2024, *K-Nearest Neighbor (KNN) Algorithm*, GeeksforGeeks, viewed 27th September 2024, <https://www.geeksforgeeks.org/k-nearest-neighbours/>.

Hassauin, S 2023, *Decision Boundary in KNN*, Kaggle, viewed 27th September 2024, <https://www.kaggle.com/code/saquib7hussain/decision-boundary-in-knn>.

Hayes, A 2024, *Autoregressive Integrated Moving Average Prediction Model*, Investopedia, viewed 26th September 2024, < https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp>.

HealthStats NSW 2024, *Explore NSW population health indicators by topic*, HealthStats NSW, viewed 17 September 2024, <https://www.airquality.nsw.gov.au/air-quality-data-services/data-download-facility>.

IMSL 2021, *What Is a Regression Model?,* IMSL by Perforce, viewed September 21 2024, <https://www.imsl.com/blog/what-is-regression-model>

Jakevdp 2024, *In Depth: k-means Clustering*, Jakevdp, viewed 22 September 2024, <https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html>.

Kaloyanova, E 2024, *How to Combine PCA and K-means Clustering in Python?*, 365 Data Science, viewed 24 September 2024, <https://365datascience.com/tutorials/python-tutorials/pca-k-means/>.

Krstic, G., Kstic, N.S. & Zambrano-Bigiarini, M. 2016, *The br2 - weighting method for estimating the effects of air pollutants on population health,* Journal of Modern Applied Statistical Methods, vol.15, no.2, pp.722-736, viewed 26 September 2024, <https://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=2042&context=jmasm>.

Messenger, G 2024, *K-Means Clustering on PCA-Transformed Ecological Data (Python, scikit-learn)*, Medium, viewed 24 September 2024, <https://medium.com/@messenger_g/k-means-clustering-on-pca-transformed-ecological-data-python-scikit-learn-9e982a1a2b15>.

Nayseem, I 2024, *Clustering with Confidence: A Practical Guide to Data Clustering in Python*, Medium, viewed September 22 2024, <https://medium.com/@nomannayeem/clustering-with-confidence-a-practical-guide-to-data-clustering-in-python-15d82d8a7bfb>.

NSW Air Quality 2024, *Data Download Facility*, NSW Air Quality, viewed 17 September 2024, <https://www.airquality.nsw.gov.au/air-quality-data-services/data-download-facility>.

Scikit Learn 2024, *2.3.11.6. Calinski-Harabasz Index*, Scikit Learn, viewed 27 September 2024, <https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index>.

Scikit Learn 2024, *2.3.6. Hierarchical clustering*, Scikit Learn, viewed September 22 2024, <https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>.

Scikit Learn 2024, *6.4.7. Estimators that handle NaN values*, Scikit Learn, viewed 20 September, 2024, <https://scikit-learn.org/stable/modules/impute.html#estimators-that-handle-nan-values>.

Scikit Learn 2024, *cross_val_score*, Scikit Learn, viewed 27 September 2024, <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html>.

Scikit Learn 2024, *Demo of DBSCAN clustering algorithm*, Scikit Learn, viewed September 23 2024, <https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py>.

Scikit Learn 2024, *Linear Regression Example*, Scikit Learn, viewed September 20 2024, <https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html>.

Scikit Learn 2024, *LinearRegression*, Scikit Learn, viewed September 20 2024, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>.

Scikit Learn 2024, *Plot Hierarchical Clustering Dendrogram*, Scikit Learn, viewed September 22 2024, <https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py>.

Scikit Learn 2024, *RandomForestRegressor*, Scikit Learn, viewed September 21 2024, <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>.

Shafi, A 2023, *K-Nearest Neighbors (KNN) Classification with scikit-learn*, DataCamp, viewed 27 September 2024, <https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn>.

StatsNotebook 2020, *Residual Plots and Assumption Checking*, StatsNotebook, viewed 28th September 2024, <https://statsnotebook.io/blog/analysis/linearity_homoscedasticity/>.

Verma, Y 2024, *Quick Way to Find p, d and q values for ARIMA*, AIM, viewed 26th September 2024, <https://analyticsindiamag.com/ai-mysteries/quick-way-to-find-p-d-and-q-values-for-arima/>.

Zvornicanin, E 2024, *Choosing the best q and p from ACF and PACF plots in ARMA-type modeling*, Baeldung, viewed 26th September 2024, <https://www.baeldung.com/cs/acf-pacf-plots-arma-modeling>.

Appendix

Appendix A: Additional Machine Learning Model: ARIMA Time Series

Another model we explored was the ARIMA (AutoRegressive Integrated Moving Average), which is widely used for time-series forecasting and is well-suited for data with temporal structures, such as the year-month format of our dataset. ARIMA is particularly effective at capturing both trends and seasonality in time-series data, even with smaller datasets like ours. The model combines autoregressive terms, which account for past values influencing future predictions, with moving average terms, which adjust the model based on the error of previous forecasts, and an integrated component to handle non-stationary data by differencing. This makes ARIMA robust for both short-term and long-term pattern recognition.

We initially considered implementing ARIMA as a core model in our project, but due to its complexity and the decision to prioritise Linear Regression as the main focus, ARIMA was sidelined. Despite this, we still included ARIMA as a secondary model to explore time-series predictions and compare its performance to other models. Specifically, ARRIMA was used to forecast Pollution and Asthma EDP levels.

One of the reasons ARIMA performed reasonably well when implemented was that we applied the Augmented Dickey-Fuller (ADF) test to ensure stationarity, which is crucial for reliable ARIMA modelling. After transforming the data to meet stationarity requirements. We used Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) plots to determine the optimal p and q hyperparameters for the model.

To assess the accuracy of the ARIMA predictions, we visually inspected the forecasted values, which closely aligned with the actual data trends. We then evaluated the model using standard error metrics, such as Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. While the predictions appeared accurate when plotted, we encountered a substantial challenge in obtaining consistently reliable values for these evaluation metrics, likely due to the small dataset size and potential overfitting issues. This difficulty in evaluating the model quantitatively was one of the reasons ARIMA was not as thoroughly refined or emphasised in comparison to the other models.