BATCH NUMBER	DB13
TEAM MEMBERS	GANTELA KIRANTEJA(20471A05K7) SHAIK SAI TEJA (21475A0508) RUDRAPOGU HEMANTH (20475A05M6)
GUIDE	CH.RAJANI
TITLE	AIR QUALITY DETECTION
DOMAIN/TECHNOLOGY	MACHINE LEARNING
BASE PAPER LINK	https://link.springer.com/article/10.1007/s13762-022-04241-5
DATASET LINK	https://www.kaggle.com/code/sharmamanali/air-quality-index-analysis-ml-visualisation

SOFTWARE REQUIREMENTS	Browser: Any latest browser like Chrome Operating System: Windows 10 Server or later Python (COLAB)
HARDWARE REQUIREMENTS	Processor: Intel® Dual Core 2.0GHz minimum Hard Disk: 1TB minimum RAM: 8GB or more
ABSTRACT	The survival of mankind cannot be imagined without air. Consistent developments in almost all realms of modern human society affected the health of the air adversely. Daily industrial, transport, and domestic activities are stirring hazardous pollutants in our environment. Monitoring and predicting air quality have become essentially important in this era, especially in developing countries like India. In contrast to the traditional methods, the prediction technologies based on machine learning techniques are proved to be the most efficient tools to study such modern hazards. The present work investigates six years of air pollution data from 23 Indian cities for air quality analysis and prediction. The dataset is well preprocessed and key features are selected through the correlation analysis. An exploratory data analysis is exercised to develop insights into various hidden patterns in the dataset and pollutants directly affecting the air quality index are identified. A significant fall in almost all pollutants is observed in the pandemic year, 2020. The data imbalance problem is solved with a resampling technique and five machine learning models are employed to predict air quality. The results of these models are compared with the standard metrics. The *Gaussian Naive Bayes* model achieves the highest accuracy while the *Support Vector Machine* model exhibits the lowest accuracy. The performances of these models are evaluated and compared through established performance parameters. The model performed the best among the other models and gets the highest linearity between the predicted and actual data.