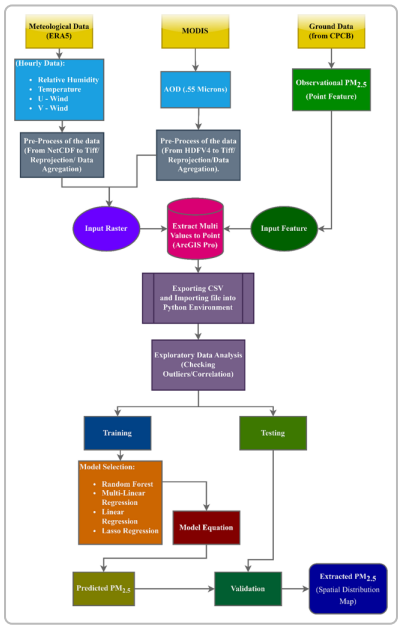
ChatGPT

Project Ideas Using ML/DSP for Maharashtra/ India Context

Air Quality ($PM_{2.5}$) Prediction in Maharashtra



A recent study in Maharashtra integrated **ground $PM_{2.5}$ measurements (CPCB stations)**, **satellite aerosol data (MODIS Fine AOD)** and **meteorological features (ERA5 winds, humidity, temperature)** to train simple ML regressors [1]. In that work, Random Forest and linear models were used to predict daily $PM_{2.5}$ (2023 data), achieving $R^2 \approx 0.87$ on held-out data [2]. For a student project, one could similarly collect CPCB hourly $PM_{2.5}$ data (publicly released for monitoring stations) and MERGED satellite AOD (via Google Earth Engine or NASA archives), join with local weather data, and train an ML model in Python. Feature engineering would include basic statistics (averages, lags) of $PM_{2.5}$ and meteorology. Kunjir *et al.* (2025) show this "data fusion" pipeline and achieved ≈95% predictions within a factor of observations [2] [1]; a student could replicate a subset of this by using RF or gradient boosting in scikit-learn.

Water Quality Prediction in Maharashtra

Maharashtra's **groundwater quality** has also been modeled by ML. The Indian National Water Monitoring Program (NWMP) publishes detailed water-quality reports (pH, dissolved oxygen, BOD, nitrates, coliform, etc.) for Maharashtra (2012–2022) [3] [4]. A recent study compiled this decade of data to predict *Water Quality Index* (WQI) or classify Water Quality Classes (good/poor) with standard regressors/classifiers. For example, Decision Trees and Polynomial Regression achieved very high accuracy ($R^2 \approx 0.99$ for WQI; ~98% accuracy for classification) [5]. In practice, one could scrape NWMP PDFs (or use data.gov.in if available), clean and standardize parameters, then train models (e.g. Random Forest, SVM, linear regression) to estimate WQI or pollution categories. This project reinforces classic ML workflows (feature scaling, train/test split) with a locally-relevant dataset.

Crop Recommendation / Yield Prediction in Maharashtra

Agriculture is vital in Maharashtra. One student project could use historical weather (rainfall, temperature from IMD or Kaggle) and soil data to **recommend crops or predict yields**. Mahale *et al.* (2024) built a crop-recommendation system using IMD climate data (2001–2022) [6]. They used Random Forest on weather features (plus an EM-based data fill) and reported ~92% accuracy in classifying the "best" crop for each division [6] [7]. A simpler project: use county-level rainfall/temperature (downloadable from IMD or Kaggle) plus known crop calendars to train a model that predicts the most suitable crop (e.g. rice, millet, cotton) for given weather profiles. Even without LSTM/deep learning, one can apply RF or Logistic Regression. Public datasets (e.g. **Agroclimatic Division** data or Kaggle's "Crop Recommendation" sets) provide example features and labels. This builds skills in supervised ML and feature engineering (e.g. compute seasonal averages, soil moisture index).

Plant Disease Classification (Coconut)

A dataset of **coconut tree disease images** is publicly available (5798 leaf images labeled into 5 diseases: bud rot, leaf spot, stem bleeding, etc.) [8]. For instance, the "Coconut Tree Disease Dataset" from 2023 provides thousands of labeled photos [8]. One project could extract visual features (color histograms, texture via GLCM, leaf shape metrics) and train an SVM or Random Forest to classify disease type. Alternatively, a student can fine-tune a lightweight CNN (e.g. MobileNet) on this data. (For reference, a recent *DeepSeqCoco* model achieved ~99.5% accuracy on a coconut disease dataset [9], though a simpler model can suffice for a class project.) This task covers DSP/image preprocessing (crop and resize leaves, augment), then ML classification. Using Python+Keras (or scikit-learn) one can build a working classifier, demonstrating basic ML on agriculture data.

Each project above uses *open datasets* (e.g. CPCB/NWMP reports, IMD/Kaggle, or public image repositories) and **clear methodology** from the literature [2] [6] [8]. The implementations can remain relatively simple (no deep network from scratch), focusing on data processing and feature-based ML models that have been shown to work in peer-reviewed studies. All cited studies are recent and open-access, providing concrete examples of methods and results in an India/Maharashtra context.

Sources: Relevant research papers and data sources as cited above provide datasets and step-by-step methods [2] [1] [6] [7] [8] [3] [4] [9] [5].

---

[1] [2] Assessing particulate matter (PM2.5) concentrations and variability across Maharashtra using satellite data and machine learning techniques | Discover Sustainability
https://link.springer.com/article/10.1007/s43621-025-01082-3

[3] [4] [5] ijfmr.com
https://www.ijfmr.com/papers/2024/5/28574.pdf

[6] [7] Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: a novel expectation-maximization technique | Discover Sustainability
https://link.springer.com/article/10.1007/s43621-024-00292-5

[8] Coconut Tree Disease Dataset - Mendeley Data
https://data.mendeley.com/datasets/gh56wbsnj5/1

[9] DeepSeqCoco: A Robust Mobile Friendly Deep Learning Model for Detection of Diseases in Cocos nucifera
https://arxiv.org/html/2505.10030v1