111120071

Mechanical Engineering

Kirthik B

SUMMER INTERNSHIP REPORT

PROJECT TITLE

FORECASTING POPULATION TRENDS IN PROXIMITY TO NUCLEAR POWERPLANTS USING MACHINE LEARNING TECHNIQUES

Introduction

- Nuclear power is a controversial topic, but it is also an important source of energy for many countries. As the global population continues to grow, there is a need to understand how the number of people living near nuclear power plants might change in the future.
- This information is valuable for planning and preparing for potential changes in these areas. For example, if we know that a population is expected to increase significantly near a nuclear power plant, we can take steps to ensure that there are adequate safety measures in place.

Research gap and Innovation

- Nuclear power plants are important, but they can also be dangerous. That's why scientists study them carefully. Researchers have looked at the risks of nuclear accidents and how many people live near nuclear power plants. But they haven't looked much at how the population around nuclear power plants might change in the future.
- This study is different. It uses machine learning to predict how the population around nuclear power plants might change in the future. This is a new way to study nuclear power plants, and it could help us make them safer and more sustainable.
- This study is important because it can help us plan for the future. If we know that the population around a nuclear power plant is going to grow, we can take steps to make sure that the plant is safe and that the people who live near it are protected.

Objectives

Proactive population prediction: Use machine learning to predict how the population around nuclear power plants might change in the future.

Data-driven insights: Use data from NASA's SocioEconomic Data and Applications Center (SEDAC) to understand population dynamics around nuclear power plants.

Enhanced prediction precision: Use feature engineering, machine learning model training, and hyperparameter tuning to improve the accuracy of population growth forecasts.

Applicable urban planning: Provide valuable information to policymakers and authorities involved in urban planning and policy formulation to help them make better decisions about nuclear power plants and the communities around them.

Comprehensive evaluation framework: Establish a robust evaluation framework using key metrics such as the R2 score, Mean Absolute Error, Root Mean Square Error, and Explained Variance to ensure the reliability of the predictions.

In other words, you want to:

Develop a new and accurate way to predict how the population around nuclear power plants might change in the future.

Use this information to help policymakers and authorities make better decisions about nuclear power plants and the communities around them.

Methodology

1) Problem Identification: Define the research problem of forecasting population growth around nuclear power plants using historical data and Machine Learning techniques.

2) Data Preprocessing: Clean, handle missing values, and address outliers or inconsistencies in the dataset sourced from NASA's SocioEconomic Data and Applications Center (SEDAC).

3) Exploratory Data Analysis (EDA): Conduct Exploratory Data Analysis to gain insights into

the dataset's characteristics, distributions, and potential patterns.

4)Feature Selection: Identify relevant features that contribute to population growth prediction and discard irrelevant or redundant attributes.

5)Population Prediction Model:

I Built a prediction model to forecast 2020 population using historical data. Use techniques like Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Sample Consensus (RANSAC), Gaussian Process Regression (GPR), Elastic Net, and K-Means Regression.

- Linear Regression:

Description: Predicts a continuous target variable using a linear relationship with input features.
Key Feature: Coefficients (weights) assigned to each feature.

- Ridge and Lasso Regression:

Description: Variations of linear regression with regularization to prevent overfitting.
Ridge Regression: Adds the sum of squared coefficients (L2 norm) to the cost.
Lasso Regression: Adds the sum of absolute coefficients (L1 norm).

- Decision Tree:

Description: Hierarchical structure for decision-making based on significant features.
Applicability: Suitable for both categorical and continuous features.
Advantage: Interpretability for visualizing decision paths.

7) Ensemble Learning

Ensemble techniques enhance machine learning models by combining individual models' strengths. The choice depends on your data, problem complexity, and the trade-offs between model complexity and prediction accuracy.

- Bagging (Bootstrap Aggregating):

Create multiple data subsets through bootstrapping.
Train separate models on each subset, like decision trees.
Average predictions to reduce overfitting and increase stability.

- AdaBoost (Adaptive Boosting):

Initially assign equal weights to data points.
Train models, giving more weight to misclassified points.
Adjust weights and iterate, combining predictions with weighted voting.
Adapts to difficult cases and handles class imbalance.

- Gradient Boosting:

Train an initial model and calculate residuals.
Train new models to predict residuals and refine the ensemble.
Handles complex relationships and is highly customizable.

- Model Averaging:

Train diverse models on the same data.
Generate predictions with all models for each new data point.
Average predictions to reduce overfitting and improve accuracy.

7) Dimensionality Reduction: Apply dimensionality reduction techniques such as Principal Component Analysis (PCA) to reduce the number of features while retaining important information.

8) K-Fold Cross Validation: Utilize K-Fold Cross Validation to assess the performance of the models and ensure their generalizability by splitting the dataset into training and validation subsets.

9)Hyperparameter tuning, achieved through GridSearch, is a method to optimize machine learning models. It follows a systematic process:

- Selection: Pick hyperparameters like learning rate or tree depth.
- Grid Creation: Define a grid with various values for each chosen hyperparameter.
- Cross-Validation: Divide the data into subsets (k-folds), then train and evaluate the model using different hyperparameter combos.
- Model Evaluation: Assess model performance for each combo and pick the one with the best results.
- Final Testing: Train the model with the best hyperparameters on the full dataset and evaluate it on an independent test set.
- The advantages include a systematic approach, automation, and finding the optimal hyperparameter configuration. However, it's essential to consider the computational cost, grid size, potential overfitting, and selecting the right evaluation metric based on your problem. This process saves time, reduces human bias, and enhances model performance.

10) Data Splitting: Split the dataset into training and testing sets, reserving the testing set for final model evaluation.
11) Model Evaluation: Evaluate the tuned models on the test dataset using evaluation metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Explained Variance, and R2 score.

- RMSE (Root Mean Square Error): RMSE measures the average difference between predicted and actual values in a regression problem. A lower RMSE suggests that the model's predictions are closer to the actual values. It's sensitive to outliers because it squares the errors.
- MAE (Mean Absolute Error): MAE is similar to RMSE but considers the absolute differences between predicted and actual values. It's less sensitive to outliers since it doesn't square the errors.
- Explained Variance: This metric indicates how well the model captures the variance in the data. A higher Explained Variance value (ranging from 0 to 1) means that the model's predictions explain a larger portion of the total variance in the actual data.
- R2 Score (Coefficient of Determination): R2 represents the proportion of variance in the target variable that can be predicted from the features. It's a normalized version of Explained Variance. A higher R2 score (ranging from 0 to 1) indicates a better fit, with 1 being a perfect fit.
- Interpretation: Lower RMSE and MAE values are better. Higher Explained Variance and R2 scores indicate better model performance, with R2 showing how well the model's predictions explain the variance relative to the mean of the target variable.

Results and Discussion

- Model Performance: Various regression models were evaluated using metrics like RMSE, MAE, R2 Score, and Explained Variance. Linear Regression, Ridge Regression, RANSAC, and Gaussian Process Regression consistently demonstrated strong performance.

- Test Size and Cross-Validation: Smaller test sizes (0.2 and 0.25) improved predictive accuracy with lower errors and higher R2 scores. Increasing K in K-fold cross-validation provided more stable results.

- Comparison of Models: Linear Regression was a robust choice for predicting population growth around nuclear power plants, consistently outperforming others. Decision Tree Regression and Artificial Neural Network (ANN) were competitive, while K-Nearest Neighbors (KNN) showed acceptable performance.

- Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) improved model performance by reducing noise and capturing essential dataset features.

- Hyperparameter Tuning: Fine-tuning improved models by optimizing internal settings to match data patterns.
- Evaluation Metrics: RMSE and MAE showed prediction error magnitude, with lower values indicating accuracy. R2 Score and Explained Variance highlighted how well models captured variance in the target variable.
- Practical Implications: Linear-based models offer accurate population predictions useful for urban planning, resource allocation, and policy-making.
- Further Research: Exploring additional variables like economic indicators, environmental factors, and demographic trends could enhance prediction precision.
- Limitations: The study's scope is limited to the provided dataset and population prediction. Model performance might vary in different scenarios.
- Future Exploration: While the selected models performed well, considering advanced techniques and refining models may further improve accuracy in similar tasks. In conclusion, linear-based models, particularly Linear Regression, are well-suited for predicting population growth near nuclear power plants, offering potential real-world applications in urban planning and policy development. However, consideration of external factors and further refinement is essential for accurate and robust predictions.

Relevance to Mechanical Engineering

- Urban Planning: Accurate population growth predictions are essential for designing infrastructure around nuclear power plants. Mechanical engineers must consider these predictions for resource allocation, energy demand, and safety measures.
- Structural Design: Mechanical engineers are responsible for ensuring the safety and efficiency of power plant structures. Population predictions help determine capacity requirements and safety protocols for plant expansions.
- Resource Allocation: Mechanical engineers optimize resource allocation and energy management. Predicting population growth aids in planning for energy generation and efficient resource distribution.
- Environmental Impact: With growing populations, environmental impact becomes crucial. Mechanical engineers in sustainable engineering use population insights to assess environmental effects, waste management, and emissions, guiding eco-friendly practices.
- Emergency Preparedness: Mechanical engineers responsible for nuclear plant safety consider potential accidents. Accurate population predictions inform emergency plans, evacuation procedures, and risk assessment, contributing to safer operations. Incorporating these predictions aligns with mechanical engineering's goals for safe, efficient, and sustainable energy use.

SUMMARY

The project focused on predicting population changes around nuclear power plants using machine learning. Data preparation and various prediction methods were employed, with combined techniques enhancing accuracy. Evaluation measures like RMSE and MAE validated the effectiveness of predictions, providing valuable insights for urban planning and emergency preparedness.