SUMMER INTERNSHIP REPORT

MACHINE LEARNING RESEARCH PROJECT UNDER DR.OSWALD C,COMPUTER SCIENCE AND ENGINEERING DEPARTMENT,NIT TRICHY

111120071

KIRTHIK B

MECHANICAL ENGINEERING A SECTION

PROJECT TITLE

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

KIRTHIK B

111120071

MECHANICAL ENGINEERING A

SUMMER INTERNSHIP REPORT

Abstract:

The increase in nuclear power plants across the globe has prompted the need for extensive research into their impact on the surrounding population. The importance of studying population growth around nuclear power plants lies in understanding the socio-economic and environmental effects they may have on the surrounding areas. India, being one of the prominent nuclear energy producers, houses a significant number of nuclear power plants across various regions. As of the latest data, there are 22 nuclear power plants in India, with an average annual population growth rate of approximately 3.5% around these facilities. The areas around these power plants are subject to varying levels of urbanization, industrialization, and demographic trends, making the population prediction crucial. A surge in population near these facilities can lead to increased demands for infrastructure, healthcare, and public services, potentially putting a strain on resources and necessitating adaptive strategies. By accurately forecasting population trends around nuclear power plants, policymakers and authorities can gain an insight on business opportunities for sustainable urban planning /development and public policy.

In this work, we aim to forecast the population growth around the proximity of nuclear power plants with the help of Machine Learning (ML) techniques. The dataset used in this study is acquired by NASA's SocioEconomic Data and Applications Center (SEDAC) which is publicly available. To the best of our knowledge, there is no detailed analytics available, using this data. Detailed feature engineering techniques on the dataset have been carried out with detailed training and testing on the dataset using various ML models. To improve the performance of our models, hyperparameter tuning and ensemble learning have also been carried out. Upon testing, we could observe interesting results on the evaluation parameters with R2 score around 0.93, reduced Mean Absolute Error and Root Mean Square Error and a high Explained variance on our models.

Table of Contents:

5. **Results and Discussion**
    - Model Performance Analysis
    - Impact of Test Size and K-Fold Cross Validation
    - Comparison of Model Performance
    - Role of Dimensionality Reduction and Hyperparameter Tuning
    - Interpretation of Evaluation Metrics
    - Practical Implications and Future Research
6. **Conclusion and Recommendations**
    - Summary of Findings
    - Contributions to Urban Planning and Policy
    - Scope for Further Refinement and Research
7. **Relevance to Field of Mechanical Engineering**
    - Urban Planning for Nuclear Power Plants
    - Structural Design and Safety Measures
    - Resource Allocation and Energy Management
    - Environmental Impact and Sustainability
    - Emergency Preparedness and Risk Assessment
8. **Summary**
9. **References**

Introduction:

This report focuses on predicting changes in population around nuclear power plants. Using existing machine learning techniques and historical data, we aim to anticipate how the number of people living near these plants might change in the future. This insight is valuable for planning and preparing for potential changes in these areas.

Literature Review:

In the article titled "Reactors, Residents and Risk," authored by Declan Butler and published in the journal Nature in 2011, the focus revolves around assessing potential hazards linked to nuclear accidents by analyzing the global distribution of populations residing in proximity to nuclear power plants. This article endeavors to shed light on the possible ramifications of nuclear accidents on densely inhabited areas located in close vicinity to these power plants.

The central objective of this work is to underscore the possible consequences of nuclear accidents in terms of human safety and lives. To this end, the author employs a world population analysis to identify regions where a significant populace could face risks in the event of a nuclear mishap. A visual representation in the form of a map is utilized to depict the population size residing within a 75-kilometer radius of various nuclear power plants across the globe. This map employs varying circle sizes and colors to denote population density, with larger and redder circles indicating higher population densities.

This analysis serves to emphasize the pressing need for robust safety measures, emergency preparedness, and meticulous risk assessment protocols. By identifying areas with substantial populations in close proximity to nuclear power plants, the article calls attention to the importance of effectively managing the safety of residents in case of a nuclear incident.

The significance of this article lies in its capacity to influence nuclear policy, emergency planning, and public awareness. Furthermore, it provides insights into the global distribution of nuclear power plants and the challenges associated with ensuring the safety of nearby inhabitants in the event of a nuclear mishap.

In summary, through a combination of quantitative analysis and geographical visualization, the article emphasizes the critical need for meticulous consideration of risks and the implementation

of safety measures in connection with nuclear energy facilities, particularly in regions marked by high population densities.

Research Gap and Innovation:

The research presented in this study introduces a novel approach to address a previously unexplored aspect in the context of nuclear power plants. While existing literature, as discussed, has focused on assessing risks and population density near these facilities, this study innovatively applies Machine Learning techniques to predict future population growth around nuclear power plants. By doing so, it bridges a gap in proactive planning for potential demographic changes in these areas.

Objectives:

1. **Proactive Population Prediction**: The primary objective of this research is to employ advanced Machine Learning techniques to predict population growth in the proximity of nuclear power plants. This objective breaks new ground by shifting the focus from retrospective analysis to proactive prediction.
2. **Data-driven Insights**: The study aims to leverage a dataset sourced from NASA's SocioEconomic Data and Applications Center (SEDAC), harnessing its potential for providing insights into population dynamics around nuclear power plants.
3. **Enhanced Prediction Precision**: Through meticulous feature engineering, training various ML models, and implementing techniques like hyperparameter tuning and ensemble learning, the research seeks to enhance the precision of population growth forecasts.
4. **Applicable Urban Planning**: An important objective is to provide valuable information to policymakers and authorities involved in urban planning and policy formulation. By delivering accurate population predictions, the research aims to contribute to sustainable development around nuclear power plants.
5. **Evaluation Metrics**: The research endeavors to establish a robust evaluation framework by employing key metrics such as the R2 score, Mean Absolute Error, Root Mean Square Error, and Explained Variance. These metrics contribute to the overall reliability of the predictions.

In summary, the innovation in this study lies in the proactive use of Machine Learning for predicting population growth around nuclear power plants. The objectives encompass proactive

prediction, data-driven insights, enhanced precision, actionable urban planning contributions, and a comprehensive evaluation framework.

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**: Build 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.Below this you can find explanation for each of these models.

   Linear Regression: Linear Regression is a simple yet powerful algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the features and the target. The algorithm finds the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted and actual values. It's characterized by coefficients (weights) assigned to each feature.

   Ridge and Lasso Regression: These are variations of linear regression that introduce regularization terms to the cost function. Ridge Regression adds the sum of squared coefficients (L2 norm) to the cost, while Lasso Regression adds the sum of absolute coefficients (L1 norm). Regularization helps prevent overfitting by penalizing large coefficients and encourages simpler models.

   Decision Tree: A Decision Tree is a hierarchical structure that makes decisions by recursively splitting the dataset based on the most significant features. Each internal node represents a decision based on a feature, and each leaf node represents a predicted outcome. Decision Trees

can handle both categorical and continuous features, and they're interpretable, making them suitable for visualizing decision paths.

Random Sample Consensus (RANSAC): RANSAC is an iterative algorithm used for fitting models in the presence of outliers. It randomly selects a subset of data points, fits a model, and identifies inliers that match the model within a predefined tolerance. RANSAC iterates to find the best model that fits the inliers.

Gaussian Process Regression (GPR): GPR is a probabilistic algorithm that models the target variable as a distribution of possible values. It uses a kernel function to capture the relationships between data points. GPR provides not only point predictions but also estimates of uncertainty associated with those predictions, making it useful for uncertainty quantification.

Elastic Net: Elastic Net combines L1 (Lasso) and L2 (Ridge) regularization terms. It aims to strike a balance between feature selection (L1) and handling correlated features (L2), making it effective for datasets with multicollinearity.

K-Means Regression: While K-Means is more commonly used for clustering, K-Means Regression applies the concept of K-Means to predict continuous values. It clusters data points into K groups and assigns the mean (centroid) of each cluster as the predicted value.

6) **Ensemble Learning**: Implement ensemble learning techniques like model averaging or stacking to combine the predictions of individual regression models for improved accuracy.
    i) Some of the models used are given below:

Bagging (Bootstrap Aggregating): Bagging is an ensemble technique that reduces variance and helps prevent overfitting by combining multiple models trained on different subsets of the dataset. It works by creating several subsets of the data through bootstrapping (sampling with replacement), training individual models on each subset, and then averaging their predictions.

Process:

b) Create multiple bootstrap samples by randomly selecting data points with replacement.

c) Train a separate model (often decision trees) on each bootstrap sample.

d) Predictions from individual models are averaged (for regression) or majority-voted (for classification) to form the final prediction.

Advantages:

e) Reduces overfitting: The averaging of diverse models reduces model variance and prevents overfitting.

f) Increases stability: The combination of predictions from multiple models stabilizes the overall prediction.

AdaBoost (Adaptive Boosting): AdaBoost is an iterative ensemble technique that focuses on improving model accuracy by giving more weight to misclassified data points. It sequentially trains a series of models, adjusting the data weights after each iteration to emphasize the misclassified samples.

Process:

g) Assign equal weights to all data points initially.

h) Train a weak model (often a decision stump) on the data, giving more weight to misclassified points.

i) Increase the weight of misclassified points and decrease the weight of correctly classified points.

j) Iterate, training new models with updated weights.

k) Combine models' predictions using weighted voting to form the final prediction.

Advantages:

l) Adapts to difficult cases: AdaBoost focuses on correcting errors, improving performance on hard-to-predict instances.

m) Handles class imbalance: It gives more weight to minority class samples, improving their representation.

Gradient Boosting: Gradient Boosting is another boosting algorithm that builds an ensemble of weak models in a stage-wise manner. It fits new models to the residual errors (differences between predictions and actual values) of the previous models.

Process:

n) Train an initial model (often a decision tree) on the data.
o) Calculate the residuals (errors) between the predictions and actual values.
p) Train a new model on the residuals, attempting to predict these errors.
q) Combine the predictions of the initial and new models.
r) Iterate, training models to predict residuals and gradually refining the ensemble.

Advantages:

s) Handles complex relationships: Gradient Boosting focuses on capturing patterns in residuals, allowing it to model complex relationships.
t) Highly customizable: It allows tuning of hyperparameters for improved performance.

Model Averaging: Model Averaging is a simple ensemble technique where predictions from multiple models are averaged to produce the final prediction. It's a versatile technique that works well when individual models exhibit complementary strengths and weaknesses.

Process:

u) Train multiple diverse models on the same dataset.
v) For each new data point, generate predictions using all trained models.
w) Average the predictions from all models to obtain the final ensemble prediction.

Advantages:

x) Reduces overfitting: Averaging predictions from diverse models helps mitigate the risk of overfitting.
y) Improved accuracy: Model Averaging leverages the strengths of different models for better predictions.

These ensemble techniques are designed to improve the performance and robustness of machine learning models by leveraging the strengths of multiple individual models. The choice between them depends on the nature of your data, the complexity of the problem, and the trade-offs you're willing to make between model complexity and prediction accuracy.

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

   Process of pca is given below:

   Centering the Data: PCA begins by centering the data, which involves subtracting the mean of each feature from the data points. This step ensures that the data is centered around the origin.

   Covariance Matrix Calculation: The next step is to calculate the covariance matrix of the centered data. The covariance between two features measures how they vary together. The diagonal elements of the covariance matrix represent the variance of individual features.

   Eigenvalue Decomposition: PCA aims to find the directions (principal components) along which the data varies the most. These directions are represented by the eigenvectors of the covariance matrix. Eigenvectors are orthogonal, meaning they're perpendicular to each other.

   Choosing Principal Components: The eigenvectors are ranked by their corresponding eigenvalues, which represent the amount of variance explained by each component. Principal components with higher eigenvalues explain more variance in the original data.

   Dimensionality Reduction: To reduce dimensionality, you select a subset of the top principal components based on how much variance you want to retain in the reduced data. These principal components are used to create a projection matrix.

   Projection: The original data is projected onto the lower-dimensional space defined by the selected principal components. This is done by multiplying the data matrix with the projection matrix.

   Advantages of PCA:

   Dimensionality Reduction: PCA allows you to reduce the number of features while retaining most of the important information in the data.

Noise Reduction: By focusing on the components that capture the most variance, PCA can help mitigate the impact of noisy or irrelevant features.

Visualization: When reducing data to two or three dimensions, PCA can aid in visualization, making it easier to understand and analyze complex datasets.

Compressing Data: PCA can be used for data compression, where you retain a subset of the principal components to represent the original data with minimal loss of 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**: Fine-tune the hyperparameters of the regression models using the training data to optimize their performance.Here I used Grid Search algorithm for this step.

GridSearch:

i) GridSearch involves defining a grid of hyperparameter values to explore. The algorithm then exhaustively evaluates all possible combinations of these hyperparameters using cross-validation and selects the combination that yields the best performance.

Process of GridSearch:

Select Hyperparameters to Tune: Determine which hyperparameters you want to tune. These could include parameters like learning rate, max depth of trees, regularization strength, etc.

Define Parameter Grid: Create a dictionary where keys are the hyperparameter names, and values are lists of values you want to try for each hyperparameter. This defines the grid of hyperparameter combinations.

Cross-Validation: Choose a suitable cross-validation strategy, such as k-fold cross-validation. The dataset is split into k subsets, and each combination of hyperparameters is trained and evaluated k times, each time using a different subset as the validation set.

Model Evaluation: For each combination of hyperparameters, train the model on the training subset and evaluate its performance on the validation subset.

Select Best Configuration: Calculate the average performance across all cross-validation folds for each hyperparameter combination. Select the combination that yields the best performance metric (e.g., highest accuracy, lowest error).

Final Model Training and Testing: Once the best hyperparameter configuration is determined, train the model on the entire training dataset using these hyperparameters. Then, evaluate the final model on an independent testing set to get an unbiased estimate of its performance.

Advantages of GridSearch:

Systematic Search: GridSearch methodically explores different hyperparameter combinations, ensuring that no combination is missed.

Automation: GridSearch automates the process of tuning hyperparameters, saving time and reducing human bias.

Optimal Configuration: GridSearch helps you find the best configuration that optimizes your model's performance based on the chosen evaluation metric.

Considerations:

Computational Cost: GridSearch can be computationally expensive, especially if you have a large number of hyperparameters and values to explore.

Grid Size: Be cautious of the grid size. A very large grid can lead to longer computation times.

Overfitting: Be wary of overfitting to the validation set during hyperparameter tuning. You can use techniques like nested cross-validation to mitigate this.

Evaluation Metric: Choose an appropriate evaluation metric based on your problem (accuracy, F1-score, etc.).

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.

Root Mean Square Error (RMSE): RMSE is a widely used metric to measure the average deviation of the predicted values from the actual values in a regression problem. It gives you an idea of how much the predictions differ from the actual values on average. RMSE is calculated by taking the square root of the average of the squared differences between predicted and actual values.

Formula: RMSE = sqrt((1/n) * Σ(predicted - actual)^2)

Interpretation: A lower RMSE indicates that the model's predictions are closer to the actual values. It's sensitive to outliers, as larger errors are squared before averaging.

Mean Absolute Error (MAE): MAE is another regression metric that measures the average absolute difference between predicted and actual values. It provides a similar insight to RMSE but is less sensitive to outliers since it doesn't square the errors.

Formula: MAE = (1/n) * Σ|predicted - actual|

Interpretation: Similar to RMSE, lower MAE values indicate better model performance, with a smaller average absolute difference between predicted and actual values.

Explained Variance: Explained Variance measures the proportion of the variance in the target variable that the model's predictions explain. It provides an indication of how well the model captures the variance in the data.

Formula: Explained Variance = 1 - (Var(predicted - actual) / Var(actual))

Interpretation: Explained Variance ranges from 0 to 1. A higher value indicates that the model's predictions explain a larger portion of the total variance in the actual data.

R2 Score (Coefficient of Determination): R2 score is a metric that represents the proportion of the variance in the target variable that is predictable from the independent variables (features). It's a normalized version of Explained Variance and is often used to assess the goodness of fit of a model.

Formula: R2 Score = 1 - (Σ(predicted - actual)^2 / Σ(actual - mean)^2)

Interpretation: R2 score ranges from 0 to 1. A higher R2 score indicates that a larger proportion of the variance in the target variable is captured by the model's predictions. An R2 score of 1 indicates a perfect fit, while a score of 0 indicates that the model's predictions are no better than simply using the mean of the target variable.

12) **Results Interpretation**: Interpret the evaluation results to identify the best-performing models and analyze the practical implications of the predictions.
13) **Conclusion and Recommendations**: Summarize the findings and insights obtained from the research, and provide recommendations for urban planning, development, and policy formulation based on the population growth predictions.

Results and Discussion:

Model Performance Analysis:

- The performance of various regression models was evaluated using key metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R2 Score, and Explained Variance.
- Linear Regression, Ridge Regression, RANSAC, and Gaussian Process Regression (GPR) consistently demonstrated strong performance across different test scenarios. These models exhibited low RMSE and MAE values and high R2 scores, indicating their ability to closely predict the 2020 population based on historical data.

Effect of Test Size and K-Fold Cross Validation:

- Smaller test sizes, such as 0.2 and 0.25, generally led to improved predictive accuracy, as reflected by lower RMSE and MAE values and higher R2 scores.
- Increasing the value of K in K-fold cross-validation provided more stable and consistent performance metrics. This indicates that the models' performance is reliable and less affected by variations in the training and testing splits.

Comparison of Model Performance:

- Linear Regression stood out as a robust model for predicting population growth around nuclear power plants. It consistently yielded the lowest errors and the highest R2 scores among the tested models.
- Decision Tree Regression and Artificial Neural Network (ANN) Regression demonstrated competitive performance, but their accuracy slightly lagged behind linear-based models.
- K-Nearest Neighbors (KNN) Regression showed acceptable performance, but its errors were slightly higher compared to other models.

Significance of Dimensionality Reduction and Hyperparameter Tuning:

- Dimensionality reduction techniques, such as Principal Component Analysis (PCA), likely contributed to improving model performance by reducing noise and capturing essential features of the dataset.
- Hyperparameter tuning further enhanced the models' predictive capacity by fine-tuning their internal settings to better match the underlying patterns in the data.

Importance of Evaluation Metrics:

- RMSE and MAE provided insights into the magnitude of prediction errors, with lower values indicating more accurate predictions.
- R2 Score and Explained Variance reflected how well the models captured the variance in the target variable. Higher values indicated models that effectively explained the variability in population growth.

Practical Implications and Further Research:

- Linear Regression, Ridge Regression, RANSAC, and GPR are promising models for predicting population growth around nuclear power plants. These accurate predictions could assist in urban planning, resource allocation, and policy-making.
- Further research could explore the inclusion of additional variables such as economic indicators, environmental factors, and demographic trends to enhance the precision of population predictions.

Limitations:

- The study's scope is limited to the provided dataset and the specific task of population prediction. Different datasets or scenarios might yield varying model performance.
- While the selected models exhibited strong performance, other advanced techniques could be explored to potentially improve accuracy further.

In summary, the results indicate that linear-based models, particularly Linear Regression, Ridge Regression, RANSAC, and GPR, are well-suited for predicting population growth around nuclear power plants. These models showcase their potential for real-world applications in urban planning and policy formulation, although further refinement and consideration of external factors are necessary for robust and accurate predictions.

Relevance to field of mechanical engineering:

Urban Planning for Nuclear Power Plants:

- The accurate prediction of population growth around nuclear power plants is crucial for designing and planning the infrastructure of these facilities.

- Mechanical engineers involved in the design and construction of power plants need to consider not only the technical aspects but also the potential impact on the surrounding population and environment.
- Your project's insights can guide mechanical engineers in estimating the future demands for resources, energy, and safety measures in accordance with the projected population growth.

Structural Design and Safety Measures:

- As the population around nuclear power plants increases, there might be a need for expanding the facilities or constructing new ones to meet the growing energy demands.
- Mechanical engineers are responsible for designing safe and efficient structures, systems, and equipment within these power plants.
- Accurate population growth predictions, as provided by your project, can aid in determining the capacity requirements, safety protocols, and emergency response mechanisms necessary for ensuring the structural integrity and stability of these facilities.

Resource Allocation and Energy Management:

- Mechanical engineers are often involved in optimizing resource allocation, energy management, and the overall efficiency of power generation processes.
- Anticipating population growth helps mechanical engineers estimate future energy consumption patterns, plan for sustainable energy generation, and optimize the distribution of resources to meet increasing demands.

Environmental Impact and Sustainability:

- With population growth, there is a heightened focus on the environmental impact of energy generation, including nuclear power.
- Mechanical engineers in the field of sustainable engineering can use your project's insights to assess the potential environmental effects of increased power generation, waste management, and emissions.
- This information can guide the development of strategies to mitigate negative impacts and promote environmentally friendly practices.

Emergency Preparedness and Risk Assessment:

- Mechanical engineers responsible for the safety and risk assessment of nuclear power plants need to consider potential scenarios involving accidents or incidents.

- Accurate population predictions can inform emergency preparedness plans, evacuation procedures, and risk assessment protocols.
- Your project's findings contribute to a more comprehensive understanding of potential risks and the implications for emergency response strategies.

Incorporating accurate population growth predictions into the planning, design, and operation of nuclear power plants aligns with the broader goals of mechanical engineering to ensure the safe, efficient, and sustainable use of energy resources.

Summary:

My project centered on forecasting population changes around nuclear power plants using established machine learning techniques. I began by preparing and refining the data for analysis. Employing tried-and-true machine learning methods, I made predictions based on historical population trends. Notably, I experimented with combining different prediction methods to achieve more accurate results.

To validate the accuracy of these predictions, I employed standard evaluation measures such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These tests confirmed the effectiveness of the combined prediction techniques in providing reliable population forecasts.

In essence, my project concentrated on leveraging machine learning to predict population shifts around nuclear power plants. This valuable insight aids in urban planning and emergency preparedness, contributing to more informed decision-making for the future of these areas.

References:

[1] Center for International Earth Science Information Network - CIESIN - Columbia University.

2015. Population Exposure Estimates in Proximity to Nuclear Power Plants, Country-Level Aggregates. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H41834D6. Accessed DAY MONTH YEAR.

[2] Butler, D. 2011. Reactors, Residents and Risk. Nature News 472 (7343): 400-401. https://doi.org/10.1038/472400a.

[3] Prediction of CardioVascular Disease (CVD) using Ensemble Learning Algorithms; Oswald, Gadi Jaya Sathwika, Arnab Bhattacharya,5th Joint International Conference on Data Science &
Management of Data (9th ACM IKDD CODS and 27th COMAD)

[4] Divorce Astrologer: Machine Learning based Divorce Prediction of Married Couples; C.Oswald, S Baranwal, SMSS Narayanan, A Bhattacharya ,2022 IEEE 19th India Council International Conference (INDICON), 1-6, 2022

[5] Regression analysis for prediction of residential energy consumption
Nelson Fumo, MA Rafe Biswas

[6] Regression analysis for prediction: understanding the process
Phillip B Palmer, Dennis G O'Connell
Cardiopulmonary physical therapy journal 20 (3), 23, 2009