Machine Learning Project Report : Power Consumption of Tetouan City

Student list:

Full name	Group	Section
YAGOUB Douaa Manel	03	01
BOUDAOUD Amira	02	01
MERS Wafaa	03	01

Abstract

This project focuses on leveraging machine learning algorithms to predict power consumption in Tetouan city, Morocco, using a comprehensive dataset spanning from January 1st, 2017, to December 30th, 2017. The dataset encompasses various environmental factors such as temperature, humidity, wind speed, and different diffuse flows, alongside power consumption data from three distribution zones: Quads, Smir, and Boussafou. The goal is to explore the efficacy of different machine learning techniques, including Decision Tree Learning, Random Forests, KNN, Naïve Bayes, SVM, and Artificial Neural Networks, in accurately predicting power consumption. Through comparative analysis, this project aims to identify the most suitable algorithm for predicting power consumption, considering historical data and environmental factors. The findings contribute to a deeper understanding of machine learning applications in energy consumption forecasting and provide insights relevant to urban energy management in Tetouan city and similar contexts.

Introduction

This report showcases in depth the exploration into the realm of machine learning, done for the purpose of a comparative analysis between different machine learning algorithms while working with the dataset “Power Consumption of Tetouan City” obtained from Power Consumption of Tetouan City - UCI Machine Learning Repository, which comprises a time series of power consumption data from three distinct distribution networks within Tetouan city, situated in northern Morocco.

The primary objective of this project is twofold. Firstly, we aim to leverage machine learning algorithms to derive regression models capable of accurately predicting power consumption within the Tetouan city networks. These regression models serve as vital tools for forecasting energy demand, facilitating efficient resource allocation, and aiding in infrastructure planning.

Secondly, our focus lies on conducting a detailed comparative analysis of different machine learning algorithms. By evaluating and contrasting the performance of these algorithms in the context of power consumption prediction, we could draw a conclusion on their strengths, weaknesses, and applicability within the domain. Through this comparative analysis, we aim to provide valuable insights into the suitability of various machine learning approaches for addressing similar real world problems in energy consumption forecasting.

In the following sections of this report, we delve into the methodology employed, the dataset characteristics, the implementation of machine learning algorithms, and a thorough analysis of the results obtained.

Dataset Description

Origin of the data:

The dataset used for this project is sourced from the UCI Machine Learning Repository, specifically the "Power Consumption of Tetouan City" dataset. This dataset provides detailed information on the power consumption across three different distribution networks in Tetouan, a city located in northern Morocco. The primary objective of this dataset is to facilitate regression tasks aimed at predicting power consumption based on various environmental and temporal factors.

Significant attributes:

DateTime: recorded at ten-minute intervals, this feature tracks power consumption over time, crucial for understanding temporal dynamics and periodicity in energy usage

Temperature: is a continuous variable representing weather temperature and is a significant predictor of power consumption due to its influence on air conditioning and cooling systems.

Humidity: another continuous variable, measures the moisture content in the air. Typically, higher humidity levels can influence the operation of dehumidifiers and air conditioning systems, thereby affecting power consumption.

Wind Speed: another continuous feature that captures wind velocity in meters per second, influencing perceived temperature and cooling requirements, and thus energy usage.

General Diffuse Flows and **Diffuse Flows:** measure solar radiation in watts per square meter, essential for understanding the impact of sunlight on power consumption, particularly for heating and cooling demands.

Zone 1, Zone 2, and Zone 3 Power Consumption: These are the target variables representing electricity usage in three distinct zones of Tetouan city, used to develop models predicting power consumption based on environmental and temporal features.

Summary statistics:

The dataset comprises 52,417 instances with key environmental features and power consumption targets. The average temperature in Tetouan is 18.81°C, with a standard

deviation of 5.82°C, and ranges from 3.25°C to 40.01°C. Humidity averages at 68.26%, with a standard deviation of 15.55%, varying between 11.34% and 94.80%. Wind speed averages at 1.96 m/s, with a standard deviation of 2.35 m/s, reaching up to 6.48 m/s. General diffuse flows average 182.70 W/m², with a standard deviation of 264.40 W/m², and diffuse flows average 75.03 W/m², with a standard deviation of 124.21 W/m², reflecting solar radiation levels. Power consumption in Zone 1 averages 32,344.97 kW, with a standard deviation of 7,130.56 kW, Zone 2 averages 21,042.51 kW, with a standard deviation of 5,201.47 kW, and Zone 3 averages 17,835.41 kW, with a standard deviation of 6,622.17 kW, illustrating the energy usage patterns across different distribution networks within Tetouan city.

These statistics indicate substantial variability in environmental conditions and power consumption, highlighting the dynamic nature of energy demand influenced by weather patterns.

Visualization:

To gain insights into the relationships between the features and the target variables, two sets of visualizations were created: a correlation matrix heatmap and scatter plots. These visualizations were obtained after performing feature engineering, which involved transforming some columns, such as DateTime, into new columns like minute, hour, and day, etc. The detailed explanation of how we derived these new columns will be provided in the next section.

Correlation Matrix Heatmap:

The heatmap illustrates the correlation between various features: environmental factors, time variables, and target attributes. Key observations include a strong positive correlation between temperature and power consumption in all zones, especially in Zone 3 (0.49). Humidity shows a notable negative correlation with temperature (-0.46) and a moderate negative correlation with power consumption in Zone 1 (-0.29). Wind speed has a significant negative correlation with humidity (-0.47) and a moderate positive correlation with temperature (0.48). There is a substantial positive correlation between general diffuse flows and diffuse flows (0.56). Temporal variables such as the month and quarter of the year are highly correlated (0.97), and they show moderate to strong correlations with power consumption, particularly in Zone 2 (0.66 and 0.32, respectively).

These correlations are significant as they indicate which features are likely to be influential predictors of power consumption. The strong correlations between temperature, month, and power consumption suggest that these variables could be critical inputs for a predictive

model. Conversely, features with weak or negligible correlations might contribute less to model performance and could be considered for exclusion to reduce complexity. Understanding these relationships will help in feature selection and engineering, ultimately improving the model's performance in predicting power consumption across the different zones.

Scatter Plots:

The scatter plots visualize the relationships between various environmental and temporal factors with power consumption across the three zones. Temperature shows a positive correlation with power consumption, particularly pronounced in Zone 3, indicating higher power usage at higher temperatures. Humidity has a more dispersed pattern, with a slight negative correlation observed, especially in Zone 1. Wind speed does not show a clear correlation with power consumption across any zone, suggesting it might be less significant as a predictor. General diffuse flows exhibit a strong positive correlation with power consumption, highlighting their potential importance in the model. Temporal variables such as month and quarter of the year show distinct patterns, with power consumption varying seasonally, peaking at certain months and quarters. The hour of the day shows clear cyclical patterns, reflecting daily usage cycles, while day of the year shows some seasonal trends. These scatter plots reaffirm the insights from the correlation matrix, indicating that temperature, general diffuse flows, and temporal variables are likely critical features for a predictive model of power consumption.

For detailed visualizations and interpretations, please refer to the notebook attached with this report.

Methodology

Feature Engineering and Selection Techniques:

Since we have three target features (for three zones), we will handle these zones separately for feature selection.

In the **feature engineering** process applied to the DateTime column, we've transformed this single timestamp into several more granular and informative temporal features. These include the day of the month to identify daily consumption patterns, the month to capture seasonal trends, the hour to detect hourly variations, and the minute for fine-scale analysis. Additionally, we extracted the day of the week to distinguish between weekday and weekend patterns, the quarter of the year to segment broader seasonal trends, and the day of the year for continuous trend analysis. These transformations enrich the dataset, enhancing the model's ability to learn and predict based on time-based patterns. The original DateTime column is then dropped to avoid redundancy.

For **feature selectio**n, we employed three techniques. Recursive Feature Elimination (RFE) is a backward elimination method where features are recursively removed based on model weights. RFE identified Temperature, Humidity, Hour, and Day of Year as critical features across all zones, highlighting their direct influence on energy needs. Random Forest Feature Importance, using a RandomForestRegressor, ranked features based on their decrease in impurity. Hour and Day of Year consistently appeared as top features, indicating strong influence due to daily and seasonal patterns, with Temperature also significant. Lasso Regression (L1 Regularization) applied L1 regularization to reduce some coefficients to zero, effectively performing feature selection. Lasso highlighted Hour, Temperature, and Day of Year as significant, aligning with the intuition that time of day and year significantly affect energy usage.

Decision Tree:

A Decision Tree Regressor was trained and tested to predict energy consumption, with the dataset divided into 80% training and 20% testing sets. Hyperparameters—max_depth, min_samples_split, and min_samples_leaf— were optimized using GridSearchCV with 5-fold cross-validation to minimize MSE. After determining the optimal parameters, the model was assessed on both sets using MSE and RMSE for error quantification and R-squared for variance explanation, ensuring accuracy in continuous outcome prediction. The use of both training and test evaluations helps in diagnosing and mitigating potential overfitting or underfitting, ensuring reliable predictions of energy consumption.

Random Forest Regressors:

A Random Forest regressor was trained and tested to predict outcomes, splitting the dataset into 80% training and 20% testing. Specific hyperparameters—number of trees, maximum tree depth, and minimum samples for node splitting—were manually configured until achieving satisfactory outcomes after using GridSearchCV with 5-fold cross-validation, but the process was excessively time-consuming and did not yield significant results. The model's performance was evaluated on both sets using MSE for error quantification and R-squared for variance explanation, ensuring accurate predictions and effective generalization to new data. This approach provided a thorough assessment of the model's predictive capability and reliability.

K-Nearest Neighbors (KNN) Classifier:

A K-Nearest Neighbors (KNN) classifier was utilized to predict outcomes, with the dataset split into 80% training and 20% testing sets. Initially, a simple KNN model was trained using the dataset with all its features then with the selected feature from the feature engineering step. Subsequently, a weighted KNN model was implemented, where the contribution of each neighbor to the prediction is weighted by the inverse of its distance. To enhance the model's performance and robustness, bagging (Bootstrap Aggregating) was applied to the KNN classifier. Then GridSearchCV was employed to optimize the hyperparameters of the KNN model. Parameters such as the number of neighbors (k), distance metric, and weighting scheme were tuned to find the best combination that minimizes Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and maximizes the R-squared score (R^2) , these were the evaluation methods and performance metrics used to evaluate our model on the current dataset. Both the training and testing sets were used to evaluate the model's performance, allowing for the assessment of potential overfitting or underfitting.

Neural Networks:

A Neural Network classifier was employed to predict outcomes, utilizing a dataset divided into 70% for training and 15% for validation and 15% . Initially, a basic Neural Network model was trained using simple architecture and parameters. Following this, a more sophisticated architecture was implemented, incorporating multiple hidden layers with varying numbers of neurons to capture complex patterns in the data. To further enhance model performance, techniques such as dropout regularization were applied to prevent overfitting. Ensemble Learning was employed to get the best out of multiple neural networks.

Additionally, hyperparameter tuning was conducted to optimize parameters such as learning rate, batch size, and activation functions. Evaluation metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared score (R^2) were utilized to assess model performance on both training and testing sets. By iteratively refining the architecture and hyperparameters through experimentation and rigorous evaluation, the Neural Network model achieved superior predictive performance on the given dataset.

SVM: Support Vector Machine:

The SVM model was trained on the full dataset, which was previously split into training and testing sets. We used the SVR library to perform regression. Due to the lengthy execution time of grid search, we adopted an experimental approach by testing all kernels to identify the best one. The RBF kernel yielded the best results across all three zones. We further optimized performance by experimenting with various values for the gamma parameter. For model evaluation, we used the same metrics as mentioned in previous sections, ensuring consistency in performance assessment.

Naive Bayes Classifier:

Since Naive Bayes is a classifier and our data is continuous, we categorized the target for the three zones into low, medium, and high. We trained two variants of the Naive Bayes model. The first was the Gaussian model, applied directly to the continuous descriptive features. For each zone, we trained the model on the full dataset, then on selected features, and finally on features after feature extraction, using grid search on the var-smoothing parameter throughout. We selected the best-performing dataset for each zone and addressed data imbalance by undersampling the target classes, followed by stacking the Gaussian models to improve results.

The second variant was the Categorical Naive Bayes model, which required binning all descriptive features before training. We used grid search to find the best value for smoothing. Throughout this process, we evaluated performance using accuracy, recall, precision, F1-score, and the confusion matrix.

Random Forest Classifier:

We also trained this model on the same data used for the Gaussian Naive Bayes model. This model was trained without any specific tuning. The primary purpose was to compare the performance of the Naive Bayes models to another model on the same data, providing a benchmark for evaluating the effectiveness of Naive Bayes in our context.

For further details about any of the models, please refer to the accompanying notebook.

Results and Analysis

Present the performance evaluation of each algorithm:

Decision tree :

Decision Tree models trained on feature-selected subsets for Zone 1 and Zone 2 demonstrate superior performance compared to those trained on the full dataset, as indicated by lower cross-validation and test Root Mean Squared Errors (RMSEs) and higher R-squared values, signifying improved fit and generalization on unseen data without inducing overfitting. Conversely, for Zone 3, performance metrics between the full dataset and feature-selected subset are similar, implying that additional features in the full dataset do not significantly contribute to overfitting or noise, suggesting that feature selection complexity may not be necessary for this zone.

Recommendations: For Zone 1 and Zone 2, it is advisable to utilize models trained on feature-selected subsets due to enhanced metrics, indicating that selected features adequately capture underlying patterns without unnecessary complexity or potential noise from less impactful features. As for Zone 3, both approaches are justifiable, with nearly identical performance metrics suggesting either model could be employed. However, opting for the simpler model with fewer features may be preferable for efficiency and interpretability unless specific features in the full dataset are crucial for further insights or operational considerations.

Random forest:

In Zones 1 and 2, the feature-selected dataset outperforms the full dataset in terms of RMSE and R-squared, indicating improved generalization by eliminating noise and irrelevant information. The reduced complexity enhances the model's focus on impactful features. Conversely, for Zone 3, minimal differences between datasets suggest either dataset could be used effectively. A notable pattern across all zones is the lower training RMSE compared to testing RMSE, indicating some overfitting, mitigated by feature selection, especially in Zones 1 and 2. The small discrepancy between training and testing RMSE in Zone 3 suggests a well-fitting model with minimal overfitting and good generalization.

Recommendation: Continue using feature-selected datasets for Zones 1 and 2 due to superior generalization. For Zone 3, either dataset is suitable, but the simpler feature-selected model may be preferable for efficiency and interpretability.

K Nearest Neighbors:

The KNN models trained on feature-selected subsets for Zones 1, 2, and 3 exhibited superior

performance compared to those trained on the full dataset. This was evidenced by Mean Absolute Errors(MAEs), Root Mean Squared Errors (RMSEs) and higher R-squared values, indicating that the feature engineering step successfully selected the most predictive features for each zone which led to a better fit and an improved generalization on unseen data without overfitting. It also simplified the model to using 2 neighbors instead of 8, for the three zones. While Fine Tuning the model, we carried on using the feature-selected data, Grid Search was further applied to get the best number of neighbors which was 2 for the three zones. Also, applying distance-based weights (Weighted Knn) improved the model performance by emphasizing nearer neighbors. Ensemble methods like bagging did not affect the performance of the models that much. For simplicity, we decided to drop it. The last step was checking for the best distance metric which turned out to be the 'Manhattan distance' for the three zones with a slight change in the number of neighbors.

Recommendations: For the three zones, it is advisable to use the weighted kNN, with 2, 3, 2 neighbors respectively for zones 1, 2, 3 and Manhattan distance. These models produce the smallest and highest r2 scores for all zones. We recommend using the feature selected data for simplicity regarding the number of features and better performance as well.

Neural networks:

The neural network models were trained using a simple architecture with 2 hidden layers for the three zones. Initial experiments tested sigmoid and ReLU activation functions, with ReLU consistently outperforming sigmoid. The ReLU activation function led to better convergence and higher predictive results, as evidenced by lower Mean Absolute Errors (MAEs), Root Mean Squared Errors (RMSEs), and higher R-squared values. For instance, the R^2 score using Relu was higher compared to the one using sigmoid, indicating improved model performance and better capture of non-linear relationships in the data.

In addition to that, the neural networks demonstrated superior performance when trained on the full dataset rather than feature-selected subsets. This suggests that the additional information from the full dataset contributed to better learning and generalization capabilities, avoiding underfitting and leveraging the complex interactions between features.

When ensemble learning methods were applied to the neural network models, there was no significant improvement in performance. The ensemble methods did not notably enhance the predictive power of the models, leading us to retain the simpler single-model approach for efficiency and clarity.

Recommendations: For predicting power consumption in the three zones, it is advisable to use a neural network with two hidden layers (64 neurons between input and first hidden layer, 32 neurons for the second with a batch size of 32) and ReLU activation functions. These models achieve the best performance metrics. Given the superior performance on the full dataset, we recommend using the complete set of features for training the neural networks to fully capture the complex relationships within the data. This approach ensures the best predictive accuracy and generalization across all zones.

Support Vector Machines:

The SVM model with the RBF kernel demonstrates promising predictive performance across all zones, with R-squared values ranging from approximately 0.895 to 0.908, indicating substantial explained variance. Notably, gamma values of 1.5 and 1.9 yielded the best results. The model exhibits relatively small errors, as seen in the mean squared error (MSE) and root mean squared error (RMSE) values. However, the model's suboptimal performance on this dataset could be attributed to its inherent characteristics not aligning well with the data's structure. SVMs heavily rely on finding optimal hyperplanes or decision boundaries, which might not adequately capture the complex, nonlinear relationships present in the dataset. Despite efforts to tune hyperparameters, the SVM's decision boundaries might not align optimally with the data distribution, resulting in suboptimal predictive performance.

Naive Bayes Classifier:

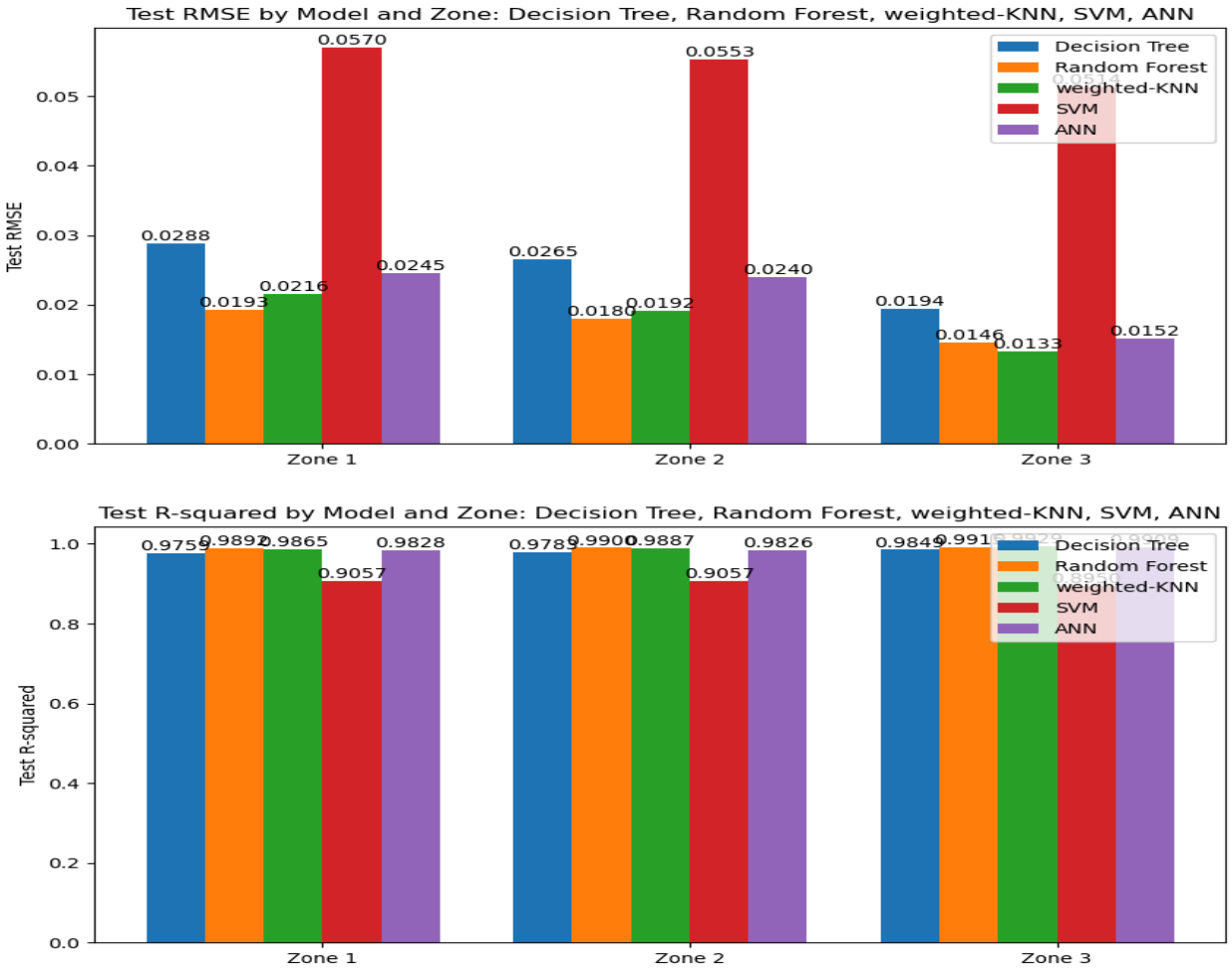
For the Gaussian Naive Bayes models, in Zone 1, the best-performing model was trained on downsampled data with stacking and smoothing values of 0.05 and 0.04, achieving an accuracy of approximately 77.06%. In Zone 2, the model trained on PCA-applied data with stacking and smoothing values of 0.01 and 0.13 achieved an accuracy of roughly 77.13%. Zone 3's optimal model utilized downsampled data with selected features and stacking with smoothing values of 0.001 and 5.0, reaching an accuracy of around 80.73%.

Across all zones, this model demonstrated consistent performance, indicating balanced performance. However, misclassifications were observed, particularly between adjacent categories.

Moving to the Categorical Naive Bayes models, in Zone 1, the model with the best smoothing parameter of 0.1 achieved the highest accuracy of approximately 85.72%. In Zone 2, the model with a smoothing parameter of 0.0 attained an accuracy of roughly 82.53%. Lastly, Zone 3 exhibited the highest accuracy of approximately 85.78% with a smoothing parameter of 0.0. For the three zones, Precision, recall, and F1-scores were also around 82% to 86%, showcasing balanced performance. Despite this, misclassifications persisted, especially between adjacent categories.

Overall, the performance of the Gaussian Naive Bayes models, while respectable, hinted at potential limitations in fitting the Gaussian distribution to all descriptive features. This constraint likely contributed to the observed misclassifications between certain target categories, suggesting that the model struggled to capture the full complexity of the data. Conversely, the Categorical Naive Bayes models demonstrated comparatively better performance. However, the process of binning the data for categorical modeling might have resulted in the loss of valuable information, thereby hindering the model's ability to make accurate predictions.

Comparative Analysis for Regression Models



The provided charts depict the performance of five different models (Decision Tree, Random Forest, weighted-KNN, SVM, ANN) in predicting power consumption across three zones (Zone 1, Zone 2, Zone 3) in Tetouan City, Morocco. The performance metrics used for comparison of the performance of these models on the **regression task** are **Root Mean Squared Error** (**RMSE**) and **R-squared (R^2)**. The following is a detailed comparative analysis of the models based on their outputs with respect to the dataset.

1. Root Mean Squared Error (RMSE):

RMSE measures the average magnitude of the errors between predicted and actual values. Lower values indicate better model performance.

Model	Zone 1	Zone 2	Zone 3
Decision Tree	0.0288	0.0265	0.0194
Random Forest	0.0193	0.0180	0.0146
weighted-KNN	0.0216	0.0192	0.0133
SVM	0.0570	0.0553	0.0514
ANN	0.0245	0.0240	0.0152

Comparative analysis :

- **Random Forest** consistently achieves the lowest RMSE across all three zones, indicating it has the best predictive accuracy among the five models.
- **weighted-KNN** also performs well , particularly in Zone 3 where it achieves the lowest RMSE of all models.
- **SVM** performs the worst in terms of RMSE in all zones, indicating it is not well-suited for this particular regression task.
- **Decision Tree and ANN** perform moderately well, with ANN generally outperforming the Decision Tree in all zones except Zone 1.

Key Insights:

- Random Forest combines multiple decision trees to reduce overfitting, leading to better generalization and lower RMSE. It can handle non-linear relationships and interactions well.In addition to this, its ensemble approach averages out the errors of individual trees, making it robust and effective in capturing complex patterns in power consumption data.

- weighted-KNN also performs consistently well, suggesting it is a good choice for this dataset, possibly due to its ability to weigh neighbors, this model also excels because it considers the proximity of historical data points with similar environmental conditions, weighing closer ones more heavily. In practical terms, if a past day had similar temperature and humidity levels, weighted-KNN can use that information to accurately predict today's power consumption.

For instance, on a hot, humid day with moderate wind, weighted-KNN can effectively predict higher power consumption due to increased air conditioning use, especially in zones with consistent patterns like Zone 3.

- SVM's poor performance could be due to the non-linear and complex nature of power consumption data, which depends on varying environmental conditions. For example, if there are intricate interactions between temperature, humidity, and wind speed that influence

power consumption, SVM might fail to model these relationships accurately, leading to less reliable predictions.

- **Decision trees** are simple and interpretable models that can capture nonlinear relationships.However, while effective, single decision trees can overfit and are less robust compared to ensembles like Random Forest, leading to higher RMSE.

For example, a Decision Tree might predict power consumption well under certain temperature and humidity conditions but fail when these conditions change slightly.

- **ANNs** are capable of capturing complex, non-linear relationships in the data through multiple layers and neurons. Also, their performance depends heavily on the architecture and training process .For instance it allows them to capture intricate patterns in power consumption data influenced by temperature, humidity, wind speed, and solar radiation.

2. R-squared (R^2):

R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables. Higher values indicate better model performance.

Model	Zone 1	Zone 2	Zone 3
Decision Tree	0.9759	0.9788	0.9849
Random Forest	0.9892	0.9900	0.9911
weighted-KNN	0.9865	0.9887	0.9929
SVM	0.9057	0.9057	0.8949
ANN	0.9828	0.9826	0.9909

Comparative analysis :

- **Random Forest** has an R^2 of 0.9892, 0.9900, 0.9911 (the highest ones) respectively for all zones. indicating it explains 99% of the variance in power consumption data confirming its superior predictive capability.
- **weighted-KNN** also shows high R^2 values, especially in Zone 3 where it outperforms even the Random Forest model slightly.
- **SVM** has the lowest R^2 values, reinforcing its poor performance observed with RMSE.
- **Decision Tree and ANN** perform reasonably well, with ANN showing slightly higher R^2 values than the Decision Tree.

Key Insights:

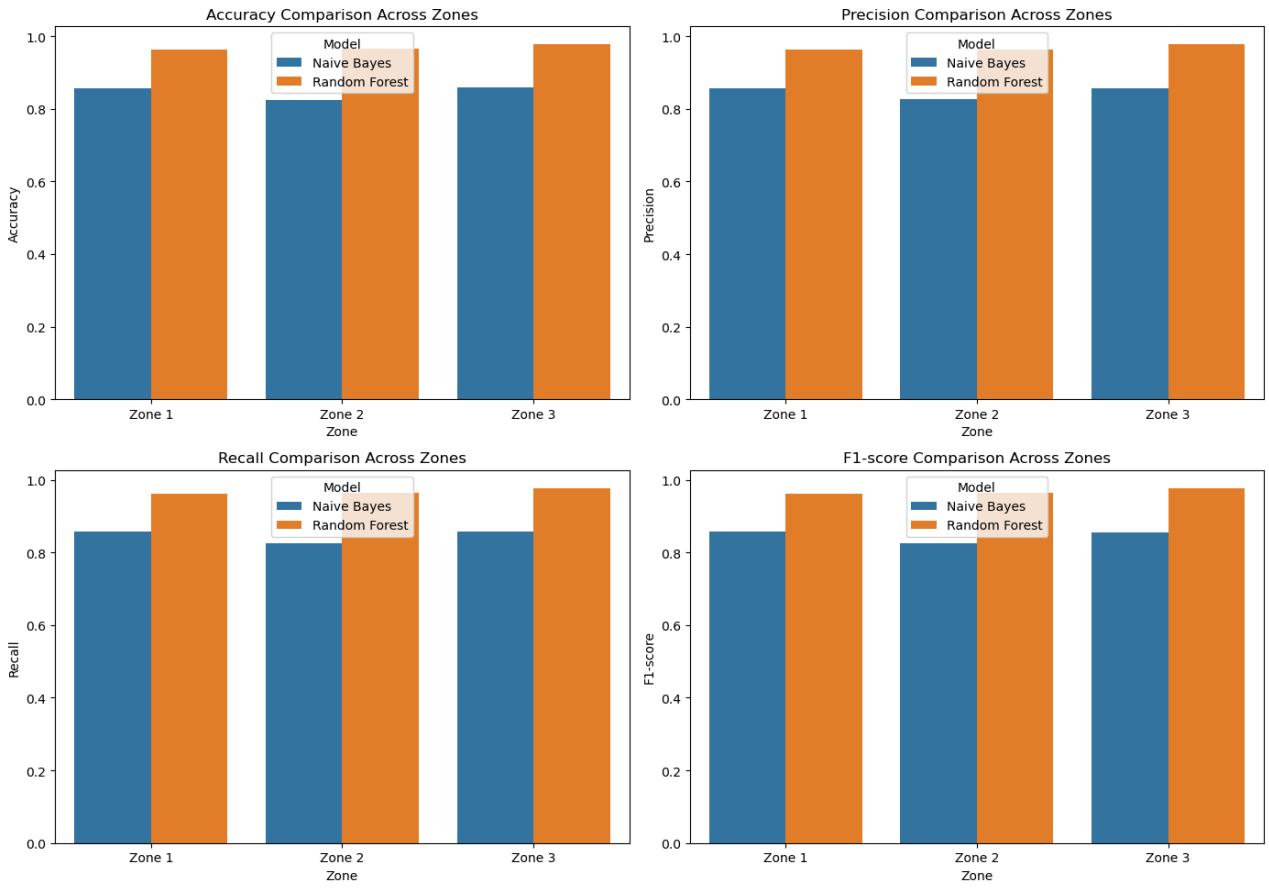
- The high R^2 value reflects the model's ability to capture the underlying patterns and relationships in the data effectively. By averaging the results of numerous trees, Random Forest reduces variance and improves generalization, capturing a wide range of patterns in the power consumption data It also shows us robustness and accuracy and consistency in performance in the three zones.
- Weighted KNN has an R^2 score which is slightly less than Random Forest, yet still very high, indicating strong predictive capabilities on our dataset. Since KNN excels in scenarios where local patterns are crucial and the weighting mechanism allows it to adapt finely to local variations, making it particularly effective for datasets where local trends influence the output significantly, this was specially noticed in the fact that KNN worked best on a different subset of features than Random Forest.
Regarding zone 3, weighted-KNN was slightly better at capturing specific patterns or local variations of this zone's power consumption data compared to the best in the model so far, Random Forest.
- Having a regression task, SVR tries to fit the data within a certain margin while aiming to find a hyperplane that best separates the data. The consistently low R^2 values (e.g., 0.8949 in Zone 3) suggest that SVM struggles to model the complex, non-linear patterns in power consumption data even after many changes of parameters and kernel choices.
- ANNs consist of multiple layers of interconnected neurons that can capture complex, non-linear relationships in data through a process of learning via backpropagation. On the other hand, they require significant tuning of hyperparameters and a large amount of data to avoid overfitting and underfitting. Training can be computationally intensive and it is advisable to choose a less complex model which requires less training, if results coincide with ANN ones.
Slightly higher R^2 values than Decision Trees (e.g., 0.9909 in Zone 3) suggest that ANN can capture more complex relationships in the data, though it may still be outperformed by the more robust ensemble methods like Random Forest.

Conclusions :

The comparative analysis underscores the strengths and weaknesses of each model in the context of predicting power consumption, considering real-life environmental conditions.
Random Forest and weighted-KNN stand out due to their robustness and ability to capture complex and local patterns, respectively. They can predict power consumption accurately by understanding the interplay of temperature, humidity, wind speed, and solar radiation.
SVM struggles due to its sensitivity to parameter tuning and inability to handle non-linearities effectively without the right settings. It may not capture the complexities of how environmental factors influence power usage.
Decision Tree and ANN offer intermediate performance, with ANN having an edge due to its capacity to model non-linear relationships, though requiring more data and careful tuning.

Zone-specific performance highlights that some areas (like Zone 3) have more predictable patterns, making it easier for models to achieve higher accuracy. This could be due to more stable environmental conditions or more consistent user behavior in that zone.

Comparative Analysis for Classification Models



The plot provided compares the performance of the best Naive Bayes model and a Random Forest model across the three zones. The comparison is made using four evaluation metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**.

Evaluation Metrics:

- **Accuracy**: Measures the proportion of correctly predicted instances out of the total instances.
- **Precision**: Indicates the proportion of true positive predictions out of all positive predictions. It measures the accuracy of the positive predictions.
- **Recall**: Represents the proportion of true positive predictions out of all actual positive instances. It measures the model's ability to capture all relevant instances.

- F1-score: Harmonic mean of Precision and Recall, providing a single measure that balances both the concerns of precision and recall.

Performance Metrics for Both Models Across Zones:

The following table summarizes the performance of the Naive Bayes and Random Forest models for each zone across the four evaluation metrics:

Metric	Zone 1 (NB)	Zone 1 (RF)	Zone 2 (NB)	Zone 2 (RF)	Zone 3 (NB)	Zone 3 (RF)
Accuracy	85.72%	96.24%	82.52%	96.45%	85.77%	97.77%
Precision	85.75%	96.24%	82.63%	96.44%	85.53%	97.80%
Recall	85.72%	96.24%	82.52%	96.51%	85.77%	97.77%
F1-score	85.64%	96.24%	82.55%	96.44%	85.49%	97.78%

Comparative Analysis

Random Forest consistently outperformed Naive Bayes in Zones 1, 2 and 3 across all metrics, indicating its robustness and reliability in these areas. Random Forest proved to be a more balanced and effective model, excelling in all evaluation metrics.

Key Insights

The superior performance of random forest can be attributed to Random Forest's ability to handle complex, non-linear relationships and its robustness against overfitting through ensemble learning. In contrast, Naive Bayes, with its assumption of feature independence, struggled to capture the intricate patterns in the data, leading to its consistent underperformance compared to the more sophisticated Random Forest model.

Discussion

Interpretation of Results

The analysis demonstrates that the Random Forest model consistently outperforms others, showcasing its robust predictive capability across all zones. With its superior performance in both RMSE and R^2 values, Random Forest emerges as the preferred choice for predicting power consumption in Tetouan City. Additionally, weighted-KNN proves effective, especially in capturing local patterns, making it a viable alternative, particularly for Zone 3 predictions. However, the poor performance of SVM suggests it may not be suitable for this regression task due to its struggles with the data's complexity. While Decision Tree and ANN models perform moderately, ANN's slight edge in capturing complex relationships makes it a viable option, especially with ample computational resources for training.

Limitations

Several limitations affect the results. Firstly, the complexity and tuning requirements of models like Random Forest and ANN are high, necessitating careful hyperparameter adjustment. SVM's performance issues highlight its sensitivity to parameter settings. Secondly, the dataset's characteristics, including potential outliers and seasonal variations, may impact model accuracy. Finally, environmental factors not present in the training data could lead to decreased model reliability under different conditions.

Potential Improvements

Improvements can be made by exploring advanced ensemble methods like XGBoost for potentially better performance. Enhanced feature engineering, incorporating additional temporal and contextual features, could help capture more data nuances. Automated hyperparameter optimization techniques, such as Grid Search or Bayesian Optimization, should be employed to find optimal model configurations. Data augmentation techniques could increase the training data's size and diversity, helping models generalize better.

Future Work

Future work should focus on developing hybrid models that combine the strengths of different approaches, such as integrating LSTM networks with Random Forest. Real-time prediction systems that dynamically update models based on new data streams should be implemented. Scenario analysis and simulations can help understand the impact of extreme conditions on power consumption. Incorporating user behavior analysis through clustering techniques can tailor predictions to different user groups. Collaborative research with meteorological experts and urban planners will enable the collection of comprehensive data, enhancing model accuracy and robustness.Moreover, further investigation into the impact of additional environmental and socio-economic variables on power consumption could enhance model performance and provide deeper insights into consumption behavior.

Conclusion

This project undertook a comprehensive comparative analysis of various machine learning models (Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, Naive Bayes, and Neural Networks) for a regression task aimed at predicting power consumption in three zones of Tataouine City, Morocco. Through meticulous model tuning and performance evaluation, we derived valuable insights into the efficacy of each algorithm in capturing the patterns within the power consumption data.

This comparative analysis enriches the existing body of knowledge by empirically demonstrating the strengths and weaknesses of different machine learning models in a specific application context. It highlights the importance of model selection and tuning in achieving optimal predictive performance. By linking model performance to both algorithmic characteristics and data patterns, this study provides a nuanced understanding of why certain models perform better than the others, and highlights the significance of using comprehensive datasets for training models, particularly in scenarios where capturing intricate relationships is crucial for accurate predictions.

From a practical side, the results can provide a framework for predicting power consumption in urban settings, offering utility companies and policymakers robust tools to forecast demand more accurately. Such predictions can inform load management strategies, optimize resource allocation, and enhance the reliability of the power grid, particularly during peak usage periods driven by extreme weather conditions.

In conclusion, this study not only advances the practical understanding of power consumption dynamics but also contributes significantly to the methodological approaches in predictive modeling, providing a valuable reference for both academic research and practical applications in energy management.

References:

- How to Improve Naive Bayes Classification Performance? | Baeldung on Computer Science
- How to deal with Imbalanced data in classification? | by Kartik Chaudhary | Game of Bits | Medium
- Feature Selection in Machine Learning | by Diborah Kiptoon | Medium
- RandomForestRegressor — scikit-learn 1.5.0 documentation
- How to Compare Machine Learning Models and Algorithms
- 1.9. Naive Bayes — scikit-learn 1.5.0 documentation
- GaussianNB — scikit-learn 1.5.0 documentation
- CategoricalNB — scikit-learn 1.5.0 documentation
- SVR — scikit-learn 1.5.0 documentation
- https://chatgpt.com/

Who did what in the project?

Data Preprocessing and Engineering : Amira BOUDAOUD
Decision Tree and Random Forest : Amira BOUDAOUD
K Nearest Neighbors : Douaa Manel YAGOUB
Artificial Neural Networks : Douaa Manel YAGOUB
Support Vector Machine : Wafaa MERS
Naive Bayes : Wafaa MERS
Comparative Analysis : All the team members
Report Writing : All the team members