Machine Learning Project Report

Power Consumption of Tetouan City

Student list:

Full name	Group	Section
Sarra ARAB	**2**	**1**
Sarah MAHMOUDI	**2**	**1**
Rachel BAKHOUCHE	**5**	**2**

Abstract

In the era of the energy crisis, Power Consumption plays a critical role in a global economy due to the imbalance between energy production and demand. Machine learning models have been widely recognized as a precise and computationally effective solution in prediction, which can help energy managers to control power systems better, improve energy usage and reduce environmental damage . In this project, we have used 7 different models to estimate the power consumption of the Tetouan city in 3 different zones (thus 3 different targets): Decision Tree Learning and Random Forests, K-Nearest Neighbor (KNN), Support Vector Regression (SVR), Naïve Bayes, Artificial Neural Network (HYBRID), Long short-term memory networks (LSTM) and Bayesian Ridge. An open-source dataset is used to validate the efficiency of the models, and different performance measures are employed to evaluate the effectiveness of the models.

Table of Content :

Introduction

The increase demand of the electric power has pushed researchers to increase their attention to the tasks related to forecasting it, the prediction task plays a crucial role for the companies of the electrical power because it provides them with great vision towards the future allowing them take the right decisions the power production and flow, just like when they need to know when to boost the power for example on the occasional days and holidays where they witness a jump in the power consumption.

The field of artificial intelligence had provided great and powerful models for the task in hand (such as naive bayes, Random forests, SVM and ANN…) and since the power consumption data are usually time series data, a further analysis can be done.Accurate power consumption predictions from these models help energy providers plan when and how much electricity to generate: this means they can make sure just the right amount of energy is produced, avoiding wasteful overproduction. Also, knowing when power demand will be high allows for smart use of renewable energy sources like solar and wind power. By using renewables more effectively, we rely less on dirty fuels that harm the environment. So, these models not only make energy management more efficient but also help protect the planet by encouraging cleaner energy practices.

this report delve into the details where set of machine learning algorithms have been applied to the Power consumption of the Titouan city

Dataset Description

Origin

The dataset utilized in this project captures various environmental and power consumption metrics for Tetouan city, Morocco. The data was recorded at ten-minute intervals, providing a granular view of the city's weather conditions and power usage patterns across three distinct zones. This high-frequency data collection allows for detailed analysis and insights into the temporal dynamics of the city's environment and energy consumption.

Significant Attributes

The dataset comprises the following features:

1. DateTime: This feature records the date and time at which each data entry was made. The data is captured at ten-minute intervals, offering a detailed temporal resolution.
2. Temperature: This feature represents the weather temperature in Tetouan city at the time of recording. It is a continuous variable measured in degrees Celsius.
3. Humidity: This feature indicates the humidity level in Tetouan city, recorded as a percentage. It is a continuous variable that provides insights into the moisture content in the air.
4. Wind Speed: This feature captures the wind speed in Tetouan city, measured in meters per second (m/s). It is a continuous variable that can influence other environmental and energy consumption factors.
5. General Diffuse Flows: This feature represents a generalized measure of diffuse flows in Tetouan city. While the specific nature of these flows is not detailed, it is a continuous variable indicative of broader flow patterns or phenomena.
6. Diffuse Flows: Similar to the General Diffuse Flows, this feature represents a specific measure of diffuse flows. It is a continuous variable that might relate to specific types of movements or distributions within the city.

And the following targets:

1. Zone 1 Power Consumption: This target variable measures the power consumption in Zone 1 of Tetouan city. It is a continuous variable recorded in kilowatts (kW).
2. Zone 2 Power Consumption: Another target variable, this measures the power consumption in Zone 2 of Tetouan city. It is also a continuous variable recorded in kilowatts (kW).
3. Zone 3 Power Consumption: The final target variable measures the power consumption in Zone 3 of Tetouan city. Like the others, it is a continuous variable recorded in kilowatts (kW).

Summary Statistics

The dataset comprises several thousand entries (52417), each representing a ten-minute interval of recorded data. Below is a summary of the key statistics for each feature:

- Temperature: The temperature ranges from 3.25°C to 40.01°C, with an average of approximately 18.81°C and a standard deviation of 5.82°C.

- Humidity: Humidity levels range from 11.34% to 94.80%, with an average value of 68.26% and a standard deviation of 15.55%.
- Wind Speed: Wind speeds range from 0.05 m/s to 6.48 m/s, with an average of 1.96 m/s and a standard deviation of 2.35 m/s.
- General Diffuse Flows: This feature ranges from 0 to 1163 units, with an average value of 182.70 units and a standard deviation of 264.40 units.
- Diffuse Flows: Diffuse flows range from 0 to 936 units, with an average of 75.03 units and a standard deviation of 124.21 units.
- Zone 1 Power Consumption: Power consumption in Zone 1 ranges from 13,895.70 kW to 52,204.40 kW, with an average of 32,344.97 kW and a standard deviation of 7,130.56 kW.
- Zone 2 Power Consumption: Power consumption in Zone 2 ranges from 8,560.08 kW to 37,408.86 kW, with an average of 21,042.51 kW and a standard deviation of 5,201.47 kW.
- Zone 3 Power Consumption: Power consumption in Zone 3 ranges from 5,935.17 kW to 47,598.33 kW, with an average of 17,835.41 kW and a standard deviation of 6,622.17 kW.

Visualization

Visualizing the dataset is crucial for uncovering patterns and relationships among the features. Key visualizations include:

- Time Series Plots: These plots display the changes in Temperature, Humidity, and Wind Speed over time, helping to identify seasonal trends and anomalies. They provide a clear view of how these environmental factors can influence power consumption.
- Heatmaps: Correlation heatmaps illustrate the relationships between different environmental variables and power consumption in the three zones. This helps to pinpoint which variables are most strongly correlated and may influence energy usage.
- Histograms: Distribution plots for Temperature, Humidity, Wind Speed, and Power Consumption reveal the frequency distributions of these variables and help identify outliers. They give an overview of the data distribution and highlight any unusual patterns.

Through exploration during EDA, visualization of the dataset provides a deep understanding of the factors contributing to power consumption in Tetouan city. This understanding is essential for developing predictive models capable of optimizing energy use and enhancing energy management.

Methodology

1.Machine learning Algorithms

In this project, we aim to deepen our understanding of various machine learning algorithms by çapplying them to a specific problem and conducting a comparative analysis of their performance. The following algorithms are implemented:

- **Decision Tree Learning and Random Forests :**

  Decision Tree Regression: Decision Tree Regression constructs a tree-like model to predict continuous target variables. It recursively splits the data based on feature values, aiming to minimize variance. Each split node represents a decision based on a feature, with leaf nodes containing predicted values. While interpretable, it may overfit.

  Random Forest Regression: Random Forest Regression aggregates predictions from multiple decision trees. Each tree is trained on a random subset of data and features. By combining predictions, it mitigates overfitting and improves accuracy. It's robust and suitable for complex regression tasks.

- **K-Nearest Neighbors (KNN)**

  K-Nearest Neighbors (KNN) is a simple, non-parametric, and instance-based learning algorithm. It classifies a data point based on how its neighbors are classified. The key idea is that similar instances are likely to have similar outputs. The algorithm works as follows:
  - Store all training instances.
  - Calculate the distance between the new instance and all training instances.
  - Identify the k-nearest neighbors (instances with the smallest distance to the new data point).
  - Determine the majority class among the neighbors (for classification) or average the values (for regression which is our case here).

- **Naïve Bayes:**

  Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, which assumes that the features in a dataset are conditionally independent given the class label. This algorithm is particularly suited for classification tasks and is known for its simplicity, efficiency, and effectiveness in handling large datasets.

The following steps describes the overall functionality of Naive bayes:

1. **Training**: Calculate the prior probabilities of each class and the likelihood of each feature given the class.
2. **Prediction**: For a new instance, compute the posterior probability for each class by combining the prior probabilities with the feature likelihoods.
3. **Classification**: Assign the class with the highest posterior probability to the new instance.

The dataset at hand is fully continuous, which presents a challenge for applying Naive Bayes directly. To address this, we propose clustering the data using the K-Means algorithm to identify the main clusters, effectively creating class labels. This transforms the task from regression to classification. The steps are as follows: first, apply K-Means clustering to the continuous data to determine the clusters. The following are the steps used for clustering the data using k-mean clustering:

1. **Initialization**: Randomly select *K* initial centroids.
2. **Assignment**: Assign each data point to the nearest centroid, forming clusters.
3. **Update**: Calculate new centroids as the mean of the data points in each cluster.
4. **Iteration**: Repeat the assignment and update steps until centroids stabilize or a maximum number of iterations is reached.

In order to determine the optimal K number of clusters the elbow method has been used.

After determining the class labels, we have used them for the Naive Bayes classifier, allowing it to operate within a classification framework.

- **Bayesian Ridge** :

In addition to the Naive Bayes algorithm, Bayesian Ridge regression has also been employed within the realm of probabilistic algorithms.

Bayesian ridge is a modified version of the Ridge regression algorithm that incorporates Bayesian methods. Unlike traditional Ridge regression, which uses a fixed regularization parameter, Bayesian Ridge regression treats the regularization parameter as a random variable with a prior distribution. This allows the model to automatically determine the most suitable regularization strength based on the data.

In our approach the bayesian ridge was used along side with ANN, where we have trained the ANN on the residuals of the bayesian ridge in order correct it an enhance the predictions.

- **SVM / SVR:**

  Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. SVM aims to find the optimal hyperplane that separates data points of different classes with the maximum margin.Support Vector Regression (SVR) is a variant of SVM suitable for regression tasks.It is used to predict continuous outcomes. The key idea behind SVR is to find a function that deviates from the actual observed values by a value no greater than a specified margin (epsilon) and is as flat as possible. Key concepts:
  - Hyperplane: A decision boundary that separates different classes.
  - Margin: The distance between the hyperplane and the nearest data points from each class.
  - Support Vectors: The data points closest to the hyperplane, which influence its position and orientation.
  - Kernel Trick: Transforming the data into a higher-dimensional space to make it linearly separable (common kernels include linear, polynomial, and RBF).

- **Artificial Neural Networks:**

  Artificial Neural Networks (ANNs) are a type of machine learning model inspired by the structure of the human brain. They consist of interconnected nodes, called neurons, organized in layers.

  Each connection between neurons has an associated weight that determines the strength of the connection. During training, the network adjusts these weights based on the input data to minimize the difference between the predicted output and the actual output, using a process called backpropagation.

In our approach, we have tried to train our data on an ANN besides we have used it in order to correct the bayesian ridge where we trained ANN on the residuals resulting from the train of the bayesian ridge.

- **LSTM (Long Short-Term Memory) :**

  Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to effectively capture long-term dependencies in sequential data. Unlike traditional RNNs, LSTM networks have a more complex structure with specialized memory cells that can maintain information over extended time periods. This allows them to better handle tasks involving sequences, such as time series forecasting. Given that our dataset is a time series, LSTM would be a suitable choice for analyzing and predicting patterns within it.

2.Feature Engineering and Selection Techniques

Feature engineering and selection are crucial steps in preparing the dataset for machine learning models, and since we have used several algorithms of machine learning we have to select the most related features for each algorithm separately.

2.1 Feature Engineering

1. Normalization/Standardization: Scaling features so they have a mean of zero and a standard deviation of one (standardization) or scaling features to a range, typically [0, 1] (normalization) in our case we have opted for Standardization.
2. Handling Missing Values: Imputing missing values using mean, median, or mode. (Not needed in this project because the dataset does not have any missing values)

2.2 Feature Selection

Feature selection is a critical step in machine learning model development, aiming to identify the most relevant features that contribute significantly to the predictive performance of the model while excluding irrelevant or redundant features. In this section, we employed Sequential Feature Selection (SFS) using the SequentialFeatureSelector class from the sklearn.feature_selection module to iteratively select the optimal subset of features for each specific zone.

We adopted a forward selection approach, which starts with an empty set of features and iteratively adds one feature at a time based on their individual performance, evaluated using the negative mean squared error as the scoring metric. The process continues until no further improvement in model performance is observed.

To account for potential variations in predictive features across different zones, we conducted feature selection independently for each zone. For each zone (Zone 1, Zone 2, and Zone 3), we initiated a LinearRegression model as the base estimator and applied the forward selection process to identify the most important features.

As a result, we obtained the selected features specific to each zone. These features represent the subset of the most influential features for predicting the target variable within the context of each zone.
Summary of Selected Features

- Zone 1 : Temperature,Wind Speed, diffuse flows, DayOfWeek, Hour, IsWeekend, Month
- Zone 2 : Temperature, diffuse flows,Hour,DayOfMonth, IsWeekend, Quarter, Season
- Zone 3 : Temperature, Wind Speed,general diffuse flows, diffuse flows, Hour, Month,Season

Note that due to redundancy, we have decided to drop general diffuse flows from the zone3's features.

By conducting zone-specific feature selection, we aimed to tailor the predictive model to the unique characteristics and dynamics present in each geographical zone, thus enhancing the model's accuracy and interpretability.

3.Model Training and Evaluation Methods

Training:

We train each of the models on 3 different sets of features corresponding to their respective targets, which indeed we splitted into training (80%) and testing (20%) (X_train1_selected, X_train2_selected ,X_train3_selected)

Evaluation:

Time series Cross-Validation :

Since our data is a time series data, the order is important therefore time series cross validation has been used to test for model consistency across different folds of the time.

Silhouette:

Since we ve used clustering algorithm (k-means) in our approach we needed to assess how well the clusters are determined by the algorithm, the silhouette metric generally assesses how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

4.Performance Metrics Used:

Since we are dealing with a regression task, we have used the following metrics:

- **Mean Squared Error (MSE):** is a metric where it basically sums the squares of the residuals then divides them over the sample size, this metric is useful for the task in hand of regression, it can give an insight on the variance of the residuals, therefore it serves as a tool to assess how accurate the predictions are.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **R-squared (R^2):**

  Used to represent the proportion of variance in the dependent variable explained by the independent variables. So it gives us an insight on how well the model was able the explain the variance of the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Mean absolute error:**

  This metric is used to give us an insight on the overall error, that on average your model is making E error value for each new instance.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Results and Analysis

Present the performance evaluation of each algorithm, comparative analysis with visualizations, and key insights and conclusions derived from the results.

Decision Tree Learning and Random Forests:

We selected Decision Tree and Random Forest models for regression tasks due to their capability to handle non-linear relationships. Six regression models were instantiated, two for each zone—one for Decision Tree and one for Random Forest—and trained using respective training data and selected features. Subsequently, the models' performance was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2), offering insights into their fitting and predictive abilities.

The results demonstrate a clear performance advantage for Random Forest models over Decision Tree models. For instance, in Zone 1, the MSE for Random Forest (0.026) is nearly half that of Decision Tree (0.048), accompanied by a lower MAE (0.108 vs. 0.135) and a higher R2 score (0.974 vs. 0.951). Similarly, in Zone 2, Random Forest outperforms Decision Tree with an MSE of 0.021 compared to 0.034, an MAE of 0.094 versus 0.110, and an R2 score of 0.979 versus 0.966. Zone 3 also shows the same trend, with Random Forest achieving lower MSE (0.025 vs. 0.046), lower MAE (0.099 vs. 0.123), and a slightly higher R2 score (0.974 vs. 0.953). These numerical differences underscore the consistent superiority of Random Forest models in predictive accuracy across all zones.

Naive Bayes:

In our approach with naive bayes we have used the K-means clustering in order to get class labels to switch the task from regression to classification, this switch will provide us a different insight on the data, where we can see it more in terms of behavior (peaks: high, low and medium).

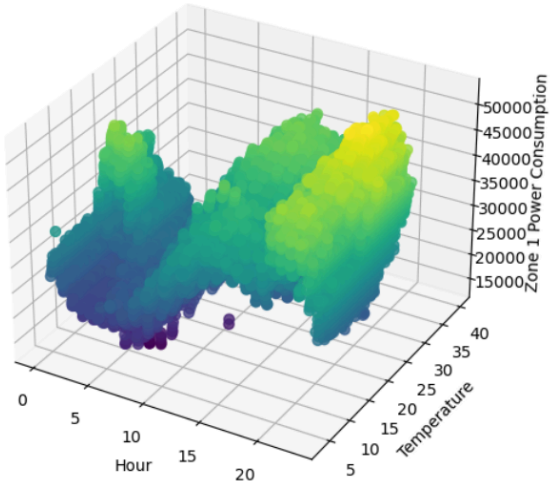
Zone1: Naive bayes with clustering:

After that we clustered the data according to the features selected for zon1, the elbow graph showed that **k=3** was the point of the elbow in the graph of the Within-Cluster Sum of Squares against the number of clusters.

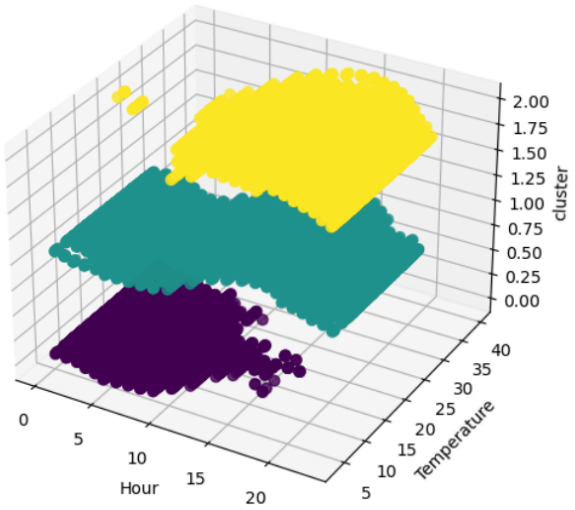
After fitting the data into k-means clusters with 3 clusters, the silhouette metric had a value of **0.253**

which signifies that the clusters are not well separated which is found to be reasonable since we are dealing with continuous features and continuous target.

After we got our labels, we fit the data into a naive bayes classifier and applied time cross validation of **ten folds**; the average accuracy across the folds was **0,97.**



figure(1)



figure(2)

This is the graph of the target Zone1 in terms of temperature and time (Hour).

Bayesian Ridge Regression Analysis : In this analysis, we applied a Bayesian Ridge Regression model, utilizing the features selected through a forward selection process. Bayesian Ridge Regression was chosen for its ability to provide probabilistic outputs and manage multicollinearity among the predictors by introducing a regularization term.

We trained the Bayesian Ridge model on the selected features, aiming to infer a linear relationship between the predictors and the target variable. The model's performance is summarized in Table 1 (see Figure 3). However, the model exhibited a relatively high error rate.

The error observed in the Bayesian Ridge model is attributed to the non-linear nature of the data, as evidenced by Figures 1 and 2. These figures illustrate the complex, non-linear relationships between the features and the target variable, which a linear model like Bayesian Ridge struggles to capture effectively.

To mitigate this issue and improve model performance, we have combined two models in order to enhance the results of the Bayesian Ridge, using ANN and Ridge.

Hybrid model : Bayesian Ridge model + ANN

In this part of our analysis, we sought to improve the predictive performance of our model by addressing the non-linear patterns in the data that the Bayesian Ridge Regression model could not capture on its own. We achieved this by training an Artificial Neural Network (ANN) on the residuals of the Bayesian Ridge model.

Step 1: Residual Analysis

Step 2: Training the ANN on Residuals

Step 3: Combining Predictions

The combined model showed a significant improvement in predictive performance. The R^2 score, which measures the proportion of variance in the dependent variable that is predictable from the independent variables, increased from 0.82 to 0.92. This increase demonstrates a substantial enhancement in the model's ability to explain the variability in the data, indicating more accurate and reliable predictions.

Application to Zones 1 and 2:

We applied the same methodology to three different zones, each representing a distinct dataset. In each zone, we observed similar behaviors and results. The combination of Bayesian Ridge Regression and the ANN consistently led to improved performance across all zones. This consistency indicates the robustness and generalizability of our approach.

Summary of Results

The results for each zone are summarized in the table below (see Figure 3). For all zones, the R^2 scores

improved significantly after applying the hybrid model, demonstrating the effectiveness of this approach in different contexts.

ANN : Zone 1:

The plot comparing the validation loss against the training loss clearly indicated overfitting. While the training loss decreased steadily, the validation loss did not follow the same pattern, showing fluctuations and eventually increasing. This divergence suggests that the model was learning the training data too well, capturing noise and details that did not generalize to the validation data.

The grid search was conducted with various combinations of parameters, including the number of neurons in each hidden layer, the learning rate, the activation functions, and dropout rates. While this approach has the potential to fine-tune the model for better performance, the extensive computational requirements limited the scope of our search and prevented us from thoroughly exploring the parameter space.

K Nearest Neighbors :

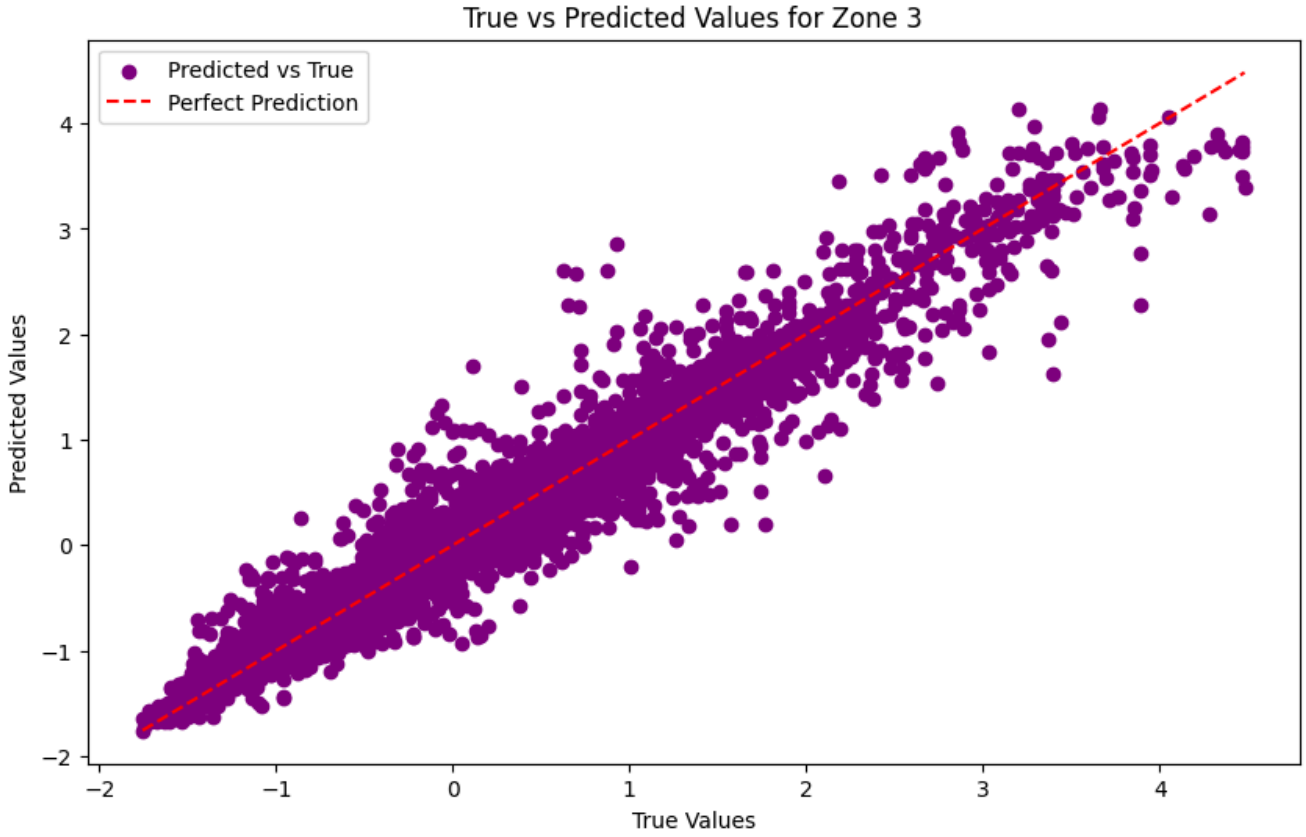
The application of K-Nearest Neighbors (KNN) regression fine tuned using gridsearch and cross-validation for predicting power consumption revealed insightful performance evaluations. For each zone, we assessed the models using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) metrics presented earlier in the report:

- Zone 1:
    - Training MSE: 0.009, Testing MSE: 0.031
    - Training MAE: 0.056, Testing MAE: 0.104
    - Training R2: 0.991, Testing R2: 0.969
- Zone 2:
    - Training MSE: 0.007, Testing MSE: 0.024
    - Training MAE: 0.048, Testing MAE: 0.092
    - Training R2: 0.993, Testing R2: 0.976
- Zone 3:
    - Training MSE: 0.027, Testing MSE: 0.042
    - Training MAE: 0.097, Testing MAE: 0.123
    - Training R2: 0.973, Testing R2: 0.957

We can deduce that there is quite a balance between test and train results where overfitting did not occur, as evidenced by the similar values of MSE, MAE, and R2 across both datasets. This balance

suggests that the models have generalized well to unseen data, maintaining high accuracy and reliability in their predictions.

The comparative analysis involved visualizing the predicted vs. true values for each zone. In Zone 3, scatter plots of the predicted vs. true values showed data points clustered around the diagonal line, representing perfect predictions. This close clustering indicated the model's high accuracy in this zone. Similar visualizations for Zones 1 and 2 also demonstrated reasonable alignment(see notebook for the figure).



To conclude, KNN is one of the best models for this task, demonstrating excellent generalization to unseen data and capturing the patterns present in the data, as indicated by the high R-squared values. This strong performance reflects the model's ability to accurately predict power consumption across different zones in Tetouane City.

Support Vector Regression (SVR) :

The application of Support Vector Regression (SVR) models for predicting power consumption revealed moderate performance evaluations. We employed three different kernels with the most common parameters: Linear SVR (kernel='linear', C=1.0, epsilon=0.1), RBF SVR (kernel='rbf', gamma='scale', C=1.0, epsilon=0.01), and Polynomial SVR (kernel='poly', degree=3, C=1.0, epsilon=0.01, coef0=1). For each zone, we assessed the models using the performance metrics presented earlier in the report and the results are presented in the table above.

For the linear kernel:

The performance of the linear kernel model seems quite consistent between the training and test sets.Considering Zone1's results: Despite being slightly less performant with an R-squared value of around 57%, the model didn't overfit, as indicated by the comparable Mean Squared Error (MSE) and Mean Absolute Error (MAE) values between the training and test sets.

The test MSE and MAE are 0.419 and 0.498, respectively, while the training MSE and MAE are 0.425 and 0.501, respectively. Additionally, the R-squared values for both the training and test sets are similar, around 0.576, indicating that the model explains approximately 57.6% of the variance in the data.

However, it's worth noting that the R-squared value of 57% suggests that the linear kernel model may not capture all the underlying patterns in the data. This indicates that there may be other factors or nonlinear relationships not captured by the linear model. Therefore, while the model's performance is consistent across training and test sets, there is room for improvement in capturing the full complexity of the data, possibly by exploring more sophisticated models.

Note that this analysis also applies for the results of Zone 2 and Zone 3.

For the RBF and polynomial kernel:

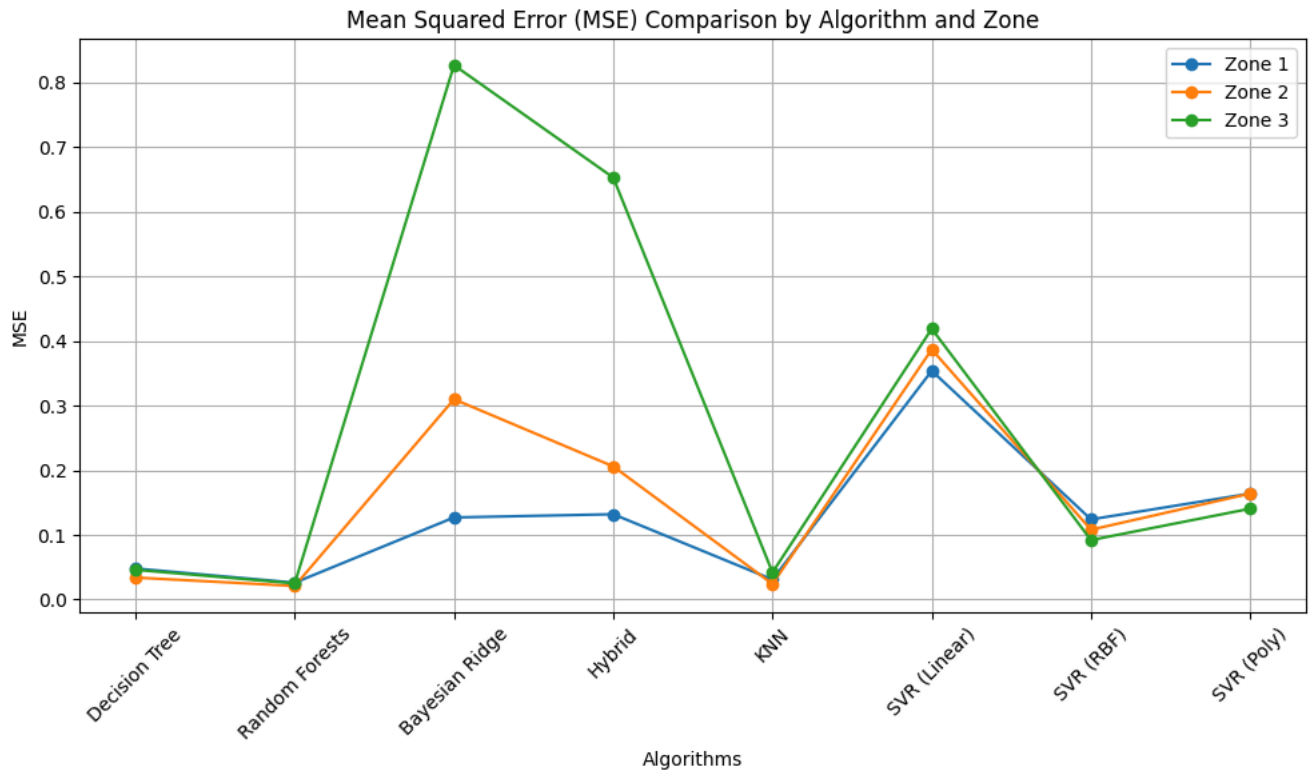
Their performance surpassed that of the linear kernel. With minimum R-squared values around 0.83 and reaching as high as 0.90, these kernels demonstrated superior predictive capabilities. This enhanced performance is due to the inherent complexity of the RBF and polynomial models, which allows them to capture a more complex understanding of the underlying data patterns.

Comparative analysis of the machine learning models :

algorithm	**Zone 1**				**Zone 2**				**Zone 3**			
metric	mse	mae	R^2	accuracy	mse	mae	R^2	accuracy	mse	mae	R^2	accuracy
Decision Tree Learning	0.048	0.135	0.951	/	0.034	0.110	0.966	/	0.046	0.123	0.953	/
Random Forests	0.026	0.108	0.974	/	0.021	0.094	0.979	/	0.025	0.099	0.974	/
Naive bayes	/	/	/	0.97	/	/	/	0.80	/	/	/	0.71

bayesian-Ridge	0.127	0.295	0.872	/	0.310	0.439	0.689	/	0.827	0.754	0.172	/
Hybrid	0.132	0.077	0.922	/	0.206	0.344	0.691	/	0.653	0.692	0.3	/
KNN	0.0309	0.104	0.969	/	0.024	0.092	0.98	/	0.042	0.123	0.96	/
SVR (linear)	0.354	0.468	0.64	/	0.387	0.487	0.613	/	0.419	0.498	0.58	/
SVR (rbf)	0.124	0.246	0.875	/	0.108	0.239	0.892	/	0.092	0.204	0.907	/
SVR (poly)	0.164	0.294	0.834	/	0.164	0.302	0.835	/	0.1406	0.272	0.859	/

figure(3): Results table



figure(4): MSE by ALgorithm & Zone

Discussion

The results of our project showcase the performance of various machine learning algorithms in predicting power consumption for different zones of Tetouan city. We observed that Random Forest models consistently outperformed Decision Tree models across all zones, demonstrating their superior predictive accuracy. Additionally, K-Nearest Neighbors (KNN) exhibited excellent generalization to unseen data with balanced performance metrics, making it a strong model for power consumption prediction tasks. Support Vector Regression (SVR) models, particularly with RBF and polynomial kernels, showed moderate performance, indicating their potential but also highlighting the complexity of the underlying data patterns.

Despite the promising results, one notable limitation is the assumption of linear relationships in certain models like Bayesian Ridge and SVR with linear kernels. While these models performed reasonably well, they may fail to capture the full complexity of the data, especially if the relationships are non-linear or involve interactions between features. Additionally, our approach of transforming the regression task into a classification task for Naive Bayes may introduce information loss and potential inaccuracies, particularly in scenarios where the data does not naturally lend itself to clustering.

To address the limitations mentioned above, several improvements can be considered. Firstly, exploring more advanced regression techniques capable of capturing non-linear relationships, such as gradient boosting machines (GBM).Furthermore, incorporating external factors such as socio-economic indicators (additional data features beyond environmental factors) or infrastructure development could enrich the predictive models, providing a more comprehensive understanding of power consumption dynamics.

Moving forward, there are several avenues for future research and development in this domain. One potential direction is the integration of real-time data streams and advanced anomaly detection techniques to improve the timeliness and accuracy of predictions. Moreover, investigating ensembles learning approaches that combine the strengths of multiple models could further enhance predictive performance for this kind of tasks (GBM mentioned earlier, AdaBoost, Bagging, Stacking…etc), Additionally, conducting comprehensive sensitivity analyses to assess the impact of different hyperparameters and model assumptions on performance could provide valuable insights for model selection and optimization.

In conclusion, our project highlights the effectiveness of some machine learning algorithms in predicting power consumption for different zones of Tetouan city. While the results showcase promising performance, there are opportunities for refinement and enhancement to further improve predictive accuracy and applicability in real-world scenarios. By addressing limitations, exploring potential improvements, and identifying possible future work, we can enhance power consumption prediction and contribute to more efficient and sustainable energy management practices.

Conclusion

In conclusion, our comprehensive analysis and evaluation of various machine learning algorithms for predicting power consumption in different zones of Tetouan city have provided valuable insights into energy management and sustainability practices.

Through the application of decision tree learning, random forests, naive Bayes, support vector regression, K-nearest neighbors, artificial neural networks, and LSTM models, we have demonstrated the diverse capabilities and performance of each algorithm in capturing the complex dynamics of the collected data. Random forest models emerged as consistently superior in predictive accuracy, while K-nearest neighbors showcased excellent generalization to unseen data.

We value the environmental enhancement that this project provides: as it empowers decision-makers with accurate predictions to optimize energy production schedules, allocate resources efficiently, and integrate renewable energy sources into the power grid. By using machine learning techniques, we can better understand and predict power consumption dynamics, making sure we're not wasting power and being kinder to the environment in Tetouan city and elsewhere.

References

1. Salam, A., & El Hibaoui, A. (2018, December). Comparison of Machine Learning Algorithms for the Power Consumption Prediction:-Case Study of Tetouan city“. In 2018 6th International Renewable and Sustainable Energy Conference (IRSEC) (pp. 1-5). IEEE.
2. https://youtu.be/c0k-YLQGKjY?si=PeFfyDVDhXTdNaEx
3. machine learning - How would you judge the performance of an LSTM for time series predictions? - Cross Validated (stackexchange.com)

Who did what in the project?

Mahmoudi sarah	Arrab Sarra	Bakhouche Rachel
Naive bayes	KNN	Decision tree/RF
Bayesian Ridge	SVR	LSTM
Hybrid model	FEATURE ENGINEERING	
ANN	EDA	