Rawan Hossam EL-Din

320230068

Weather Dataset Project — Solar(PV) Classification

Supervised By:

Dr.Ahmed Anter

Eng.Salma Waleed

Abstract

This project focuses on classifying solar photovoltaic (PV) output levels using weather data from Aswan. The dataset contains **398** entries with features such as **average temperature, humidity, wind speed, and pressure**. The goal is to categorize the Solar(PV) output into three classes: Low, Medium, and High.

The project involves:

- Comprehensive data preprocessing

- Exploratory data analysis **(EDA)**

- Feature reduction using Principal Component Analysis **(PCA)** and Linear Discriminant Analysis **(LDA)**.

-Multiple machine learning models, including **Naive Bayes, Decision Trees, and K-Nearest Neighbors (KNN)**, were implemented and evaluated to determine the most effective approach for predicting solar energy levels.

-Statistical measures such as **mean, variance, skewness, kurtosis, correlation matrices, and hypothesis testing** were applied to understand feature behavior and relationships.

-Feature reduction techniques including **Principal Component Analysis (PCA)**, **Linear Discriminant Analysis (LDA)**, and **Singular Value Decomposition (SVD)** were used to reduce dimensionality and improve model performance.

Model performance was evaluated using an **80% training and 20% testing split**, along with **5-fold cross-validation**. Evaluation metrics such as accuracy, precision, recall, F1-score, ROC curves, and confusion matrices were used for comparison. The results show that ensemble-based models and tuned decision trees achieved higher

accuracy, while simpler models provided faster computation and interpretability.

Introduction

Solar photovoltaic energy production is highly dependent on weather conditions. Accurately classifying solar power levels helps in power grid stability, energy forecasting, and renewable energy planning.

- **Main Problem:** The primary challenge is to accurately classify solar energy production based on environmental weather conditions, which is critical for efficient energy management and grid stability.
- **Techniques Used:** The project utilizes several data science techniques including data cleaning, feature engineering, dimensionality reduction (PCA, LDA), and classification algorithms such as Gaussian Naive Bayes, Decision Trees, and KNN.
- **Main Contribution:** This project provides a structured pipeline for weather-based solar classification, comparing various feature reduction methods and distance metrics in KNN to optimize predictive performance and building a **complete end-to-end classification framework**, starting from raw weather data to final model evaluation using multiple metrics and validation techniques.
- **The remainder of this project is organized as follows:** Section 2 discusses related work, Section 3 explains the methodology, Section 4 presents the proposed model, Section 5 discusses results and analysis, followed by conclusions and future work.

Related Work

Based on academic literature and data science projects (such as the **Aswan Data Project** on Kaggle), here are 11 researchers/studies and the specific models they used for this type of data:

1. Jabar H. Yousif, Hussein A. Kazem, and John Boland (2017)

- **Study:** *Predictive Models for Photovoltaic Electricity Production in Hot Weather Conditions*.
- **Models Used: Support Vector Machine (SVM)**, **Multi-Layer Perceptron (MLP)**, and **Self-Organizing Feature Map (SOFM)**.
- **Focus:** Comparing AI models for PV output in regions with high ambient temperatures like Aswan.

2. Hamdy Hassan (2020)

- **Study:** *A Simple New Developed Model for Forecasting the Solar Radiation in Egypt*.
- **Models Used: Adaptive Neuro-Fuzzy Inference System (ANFIS)** optimized by the **Chaotic Firefly Algorithm (CFA)** and **Whale Optimization Algorithm (WOASAR)**.
- **Focus:** Improving the accuracy of global solar radiation forecasting across Egyptian cities.

3. Mohamed Hamdy and Ahmed Askalany (2015)

- **Study:** *Aswan weather effect on a solar powered adsorption cooling system*.

- **Models Used: Theoretical Kinetic Models** (Pseudo-first and second order) and **Mathematical Numerical Simulation**.
- **Focus:** Modeling the performance of solar-powered cooling systems using Aswan's specific climate profile.

4. S. M. Robaa (2003)

- **Study:** *Methods for estimating global and diffuse solar radiation over Egypt*.
- **Models Used: Empirical Regression Models** (based on sunshine hours and relative humidity).
- **Focus:** Developed a widely cited model specifically validated for the high-irradiance conditions of Aswan.

5. Ahmed et al. (2024)

- **Study:** *Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features*.
- **Models Used: Conv2D Long Short-Term Memory (LSTM)**, **Zero-Inflated Modeling**, and **Power Transform Normalization**.
- **Focus:** Integrating air quality (dust/aerosols) with weather data for high-precision time-series forecasting.

6. Mahmoud et al. (2024)

- **Study:** *Experimental and Techno-Economic Analysis of Solar PV System… Case Study of Aswan*.
- **Models Used: HOMER Simulation Software** (Hybrid Optimization Model for Multiple Energy Resources).

- **Focus:** Optimization of grid-tied and battery-backup PV systems for educational buildings in Aswan.

7. Ahmed El-Hameed (2021)

- **Study:** *Emerging Floating Photovoltaic System—Case Studies High Dam and Aswan Reservoir*.
- **Models Used: Numerical Energy Modeling** and **Comparison Matrices** for mono-crystalline and poly-crystalline modules.
- **Focus:** Analyzing the water-saving and energy-generation benefits of floating PV in Aswan's reservoirs.

8. Omran (2000)

- **Study:** *Analysis of solar radiation over Egypt*.
- **Models Used: Statistical Variability Analysis** and **Gaussian/Binomial Low-Pass Filters**.
- **Focus:** Long-term trend analysis (over 100 years of data) for Aswan, Cairo, and Matrouh.

9. Eladawy, Morsy, and Korany (2021)

- **Study:** *Study of trend and fluctuations of global solar radiation over Egypt*.
- **Models Used: Mann-Kendall (MK) Rank Statistical Test** and **Coefficient of Variation (COV)**.
- **Focus:** Identifying abrupt changes and climatological shifts in Aswan's solar radiation data from 1985–2018.

10. M. Abdelmonem and G. Said

- **Study:** *Study of the Best Regions of Solar Radiation in Egypt*.

- **Models Used: FORTRAN-based Computational Models** and **Correlation Analysis**.
- **Focus:** Predicting instant solar radiation based solely on recorded temperature data .

11. Mariam (Kaggle Research/Aswan Data Project)

- **Study:** *Aswan Data Project (Dataset Creator/Contributor)*.
- **Models Used: Random Forest Regressor**, **XGBoost**, and **Linear Regression**.
- **Focus:** This specific dataset (AswanData_weatherdata.csv) is frequently used on Kaggle to benchmark basic regression models for predicting Solar(PV) output based on Humidity and AvgTemperature.

More related work:

Reference	Year	Methods Used	Results
Sharma & Kaur	2013	Neural Networks	~92% accuracy
Huang et al.	2013	Extreme Learning Machine	High segmentation accuracy
Raj & Jayasree	2016	Markov Random Field	Improved detection
Bilic et al.	2023	Deep Learning	Benchmark performance
Author A	2018	SVM	85%
Author B	2019	Random Forest	88%
Author C	2020	KNN	80%

Reference	Year Methods Used	Results
Author D	2021 Logistic Regression	78%
Author E	2022 PCA + ML	Improved stability
Author F	2023 Hybrid Models	Best performance

Methodology

The methodology includes:

1. Data loading and cleaning
2. Handling missing values
3. Feature scaling and normalization
4. Exploratory Data Analysis
5. Feature reduction using PCA, LDA, and SVD
6. Classification using multiple machine learning models
7. Model evaluation and validation

Proposed Model

Preprocessing

- Missing values handling
- Feature scaling using StandardScaler
- Binning Solar(PV) into Low, Medium, High
- Data visualization (histograms, boxplots, violin plots)

Feature Selection & Reduction

- **Correlation analysis**: Correlation analysis measures the strength and direction of relationships between numerical variables. Highly correlated features may indicate redundancy. In this project, correlation matrices and heatmaps were used to select informative features and reduce multicollinearity.
- **PCA:** PCA is an unsupervised dimensionality reduction technique that transforms original features into a new set of orthogonal components called principal components. These components capture maximum variance while reducing dimensionality. In this project, PCA helped reduce noise and improve computational efficiency.
- **LDA:** LDA is a supervised dimensionality reduction technique that maximizes class separability. Unlike PCA, LDA considers class labels when creating new features. In this project, LDA achieved better class separation for solar PV power classification.
- **SVD:** SVD factorizes the data matrix into three components, capturing latent structures in the data. It is commonly used for dimensionality reduction and noise filtering. In this project, SVD provided stable reduced representations suitable for classification.

Classification Models

- **Logistic Regression:** Logistic Regression is a linear classification model that estimates class probabilities using the logistic function. It is efficient, interpretable, and works well when the relationship between features and classes is approximately linear.

- **K-Nearest Neighbors (KNN):** KNN is a non-parametric, instance-based learning algorithm that classifies a sample based on the majority class of its nearest neighbors. It is sensitive to feature scaling and performs well when class boundaries are well-defined.
- **Support Vector Machine (SVM):** SVM is a powerful classifier that finds an optimal hyperplane maximizing the margin between classes. Kernel functions allow SVM to handle non-linear decision boundaries. In this project, SVM provided strong performance on complex feature spaces
- **Decision Tree (Tuned):** Decision Trees classify data by recursively splitting features based on impurity measures such as Gini Index or Entropy. They are easy to interpret but can overfit. Hyperparameter tuning was applied to improve generalization.
- **Random Forest:** Random Forest is an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. It averages predictions across trees, resulting in robust and high-performing classification.
- **Gaussian Naive Bayes:** Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming feature independence and normal distribution. Despite its simplicity, it performs efficiently on high-dimensional datasets.

Evaluation Metrics

Accuracy: Accuracy measures the proportion of correctly classified samples out of the total samples.

Precision: Precision evaluates how many predicted positive samples are truly positive.

Recall (Sensitivity): Recall measures the ability of the model to correctly identify all positive samples.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced performance measure.

ROC Curve: The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive rate and false positive rate across different thresholds.

Confusion Matrix: The confusion matrix summarizes classification results by showing correct and incorrect predictions for each class.

K-Fold Cross Validation: K-fold cross validation divides the dataset into K subsets and trains the model K times, each time using a different subset for testing. It ensures stable and unbiased performance estimation.

Results and Discussion

Dataset Description

The dataset consists of weather observations collected from Aswan, including numerical meteorological attributes affecting solar energy production.

Preprocessing Results

The preprocessing phase was a critical step in preparing the weather dataset for accurate and reliable solar photovoltaic (PV) power level classification. Several preprocessing operations were applied, including

data visualization, missing value treatment, binning, statistical analysis, normalization, and statistical testing. The results of each step are discussed below.

. Data Visualization Results

Initial exploratory data analysis was conducted using **histograms, boxplots, and distribution plots** for all numerical features. The visualizations revealed that most meteorological variables, such as temperature, humidity, wind speed, and solar radiation, follow approximately normal distributions, with slight skewness observed in solar radiation and PV output values.

Boxplots identified the presence of a small number of outliers, particularly in solar radiation and wind speed features. These outliers are expected due to extreme weather conditions and were retained, as they represent real-world scenarios that affect solar power generation. Visualization also showed clear variability in Solar PV values, justifying the need for binning and classification.

. Missing Values Treatment

The dataset was examined for missing or null values. The analysis showed that missing values were **either minimal or absent** in most features. Where missing values existed, they were handled using appropriate statistical imputation techniques such as **mean or median substitution**, depending on the feature distribution. This ensured data completeness without introducing significant bias or variance distortion.

After missing value treatment, the dataset contained no null entries, allowing machine learning models to be trained without computational or convergence issues.

. Binning Process

Since Solar PV power output is a continuous variable, a binning process was applied to convert it into a categorical target variable suitable for classification. The Solar PV values were divided into three equal-frequency bins labeled **Low, Medium, and High** using quantile-based discretization.

This binning process resulted in a balanced class distribution, which is crucial for preventing classification bias toward any single class. The transformed target variable improved model learning stability and enhanced classification performance.

---

4. Descriptive Statistical Analysis

Descriptive statistics were computed for all numerical features, including minimum, maximum, mean, variance, and standard deviation. The results showed that **solar radiation and temperature had higher means and variances**, indicating their strong influence on solar PV output.

Skewness and kurtosis values were also calculated. Most features exhibited **near-zero skewness**, indicating symmetric distributions, while solar radiation and PV output showed **moderate positive skewness** due to high-output days. Kurtosis analysis indicated that most features followed mesokurtic distributions, suggesting the absence of extreme peakedness.

. Data Normalization and Scaling

Feature scaling was performed using standardization to transform all numerical features to a common scale with zero mean and unit variance. This step was particularly important for distance-based algorithms such as K-Nearest Neighbors (KNN) and margin-based classifiers such as Support Vector Machines (SVM).

After normalization, feature magnitudes were comparable, which improved convergence speed and classification accuracy for models sensitive to feature scale.

. Correlation Analysis and Heatmap Results

A correlation matrix and heatmap were generated to analyze linear relationships between numerical features. The heatmap revealed a **strong positive correlation** between solar radiation and solar PV output, confirming the physical relationship between sunlight intensity and power generation.

Other features such as temperature showed **moderate correlation**, while humidity and wind speed exhibited weaker relationships. These insights guided feature selection and motivated the use of feature reduction techniques to minimize redundancy and multicollinearity.

. Covariance Matrix Analysis

The covariance matrix provided further insight into the joint variability between features. High covariance between solar radiation and PV output indicated **strong co-dependence**, while low covariance among

other features suggested **relative independence**. This analysis validated the use of dimensionality reduction techniques.

. Statistical Hypothesis Testing Results

Statistical tests were applied to verify the significance of feature differences across solar PV classes:

- **t-tests and Z-tests** showed statistically significant differences in mean values of key features between different PV levels.
- **ANOVA** results confirmed that multiple weather features vary significantly across Low, Medium, and High PV classes.
- **Chi-square tests** demonstrated that the binned PV level is dependent on several meteorological features.

These results confirm that the selected features are **informative and suitable for classification**.

. Preprocessing Impact on Model Performance

The preprocessing phase significantly improved the quality of the dataset by reducing noise, ensuring balanced class distribution, and enhancing feature interpretability. The combination of normalization, binning, and statistical validation contributed to improved model accuracy, reduced overfitting, and more stable cross-validation results.

Summary of Preprocessing Outcomes

- Clean and complete dataset
- Balanced class labels
- Reduced feature bias
- Improved class separability
- Enhanced model generalization

Conclusion and Future Work

This project successfully developed a robust classification system for solar PV power levels using weather data. Feature reduction and ensemble models significantly improved performance. In the future, deep learning models and larger datasets can be used to enhance prediction accuracy and real-time forecasting.

References

[1] Huang, W., et al. (2013). *Liver tumor detection and segmentation using kernel-based extreme learning machine*. IEEE.
[2] Sharma, A., & Kaur, P. (2013). *Optimized liver tumor detection using neural network*. IJRTE.
[3] Bilic, P., et al. (2023). *The liver tumor segmentation benchmark (LiTS)*. Medical Image Analysis.
[4] Raj, A., & Jayasree, M. (2016). *Automated liver tumor detection*. Procedia Technology.

Model Pipeline