Solar Photovoltaic Power Prediction Using Weather Data and Machine Learning Techniques

Abstract :

The rapid growth of renewable energy systems has increased the need for accurate prediction of solar photovoltaic (PV) power generation to enhance energy planning and grid stability. Solar PV output is highly dependent on meteorological conditions such as temperature, humidity, wind speed, and solar radiation, which makes data-driven modeling an effective solution for forecasting performance. This project presents an end-to-end machine learning pipeline for analyzing and predicting solar photovoltaic power generation using weather data collected from Aswan, Egypt. [8]

The proposed system begins with data preprocessing and exploratory data analysis to ensure data quality and reliability. Missing values are handled using statistical imputation methods, while descriptive statistics such as mean, variance, skewness, and kurtosis are computed to understand the distribution of weather features. Correlation and covariance analyses are performed to identify linear relationships between variables, and data visualization techniques including histograms, boxplots, and heatmaps are applied to detect outliers and patterns. Statistical hypothesis tests, including Chi-square tests, independent t-tests, and ANOVA, are used to evaluate the significance of weather features in relation to solar PV output.

Feature selection and dimensionality reduction techniques are then applied to improve model performance and reduce complexity. SelectKBest with ANOVA F-test is used to identify the most influential features. In addition, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD) are employed to reduce dimensionality while preserving important information. Several machine learning classifiers, including Gaussian Naive Bayes, Decision Tree, Linear Discriminant Analysis, and K-Nearest Neighbors with different distance metrics, are trained and evaluated.
The dataset is split into 80% training and 20% testing subsets, and model performance is assessed using accuracy, precision, recall, F1-score,

confusion matrices, and k-fold cross-validation. The results demonstrate that machine learning techniques can effectively model the relationship between weather conditions and solar PV output, with certain classifiers achieving higher predictive accuracy and better generalization. This study highlights the importance of preprocessing, feature engineering, and model comparison in developing reliable solar energy prediction systems.

Introduction :

Solar energy is one of the most promising renewable energy sources due to its sustainability, environmental benefits, and wide availability. However, the power generated by solar photovoltaic (PV) systems is highly variable and influenced by several weather-related factors, including temperature, solar radiation, wind speed, and atmospheric conditions. This variability creates challenges in energy forecasting, grid management, and efficient utilization of solar power. As a result, accurate prediction of solar PV output has become an important research topic in renewable energy and data science. [1],[3],[4].

Traditional analytical models often struggle to capture the complex and nonlinear relationships between weather variables and solar PV output. Machine learning techniques offer a powerful alternative by learning patterns directly from data without requiring explicit physical modeling. By leveraging historical weather data, machine learning models can provide accurate and adaptable predictions of solar power generation.

In this project, an end-to-end machine learning framework is developed to analyze and predict solar photovoltaic output using weather data from Aswan, Egypt. The project applies data preprocessing, statistical analysis, feature selection, dimensionality reduction, and multiple classification algorithms to model solar PV behavior. Techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD) are employed to reduce data dimensionality and enhance model performance.

The main contribution of this project is the implementation and comparison of multiple machine learning classifiers combined with statistical testing and feature reduction techniques to identify the most effective approach for solar PV prediction. Additionally, the project provides a comprehensive evaluation framework using train-test splitting, k-fold cross-validation, and confusion matrix analysis.

The remainder of this report is organized as follows. Section II reviews related work in solar PV prediction using machine learning methods. Section III describes the methodology and proposed model, including preprocessing and feature engineering techniques. Section IV presents the experimental results and discussion. Finally, Section V concludes the project and suggests directions for future work.

Related Work :

Solar photovoltaic (PV) power prediction has been extensively studied using a variety of machine learning, deep learning, and hybrid methods. Most studies report improved performance over traditional statistical models by leveraging weather and historical power data [2] ,[9] ,[5]. The following works highlight key approaches and results that are closely related to the methodology applied in this project:

Chakraborty et al. (2023) applied ensemble machine learning models (e.g., Bagging, Boosting, Stacking) to predict solar PV generation using meteorological data. Their stacking and voting ensemble models achieved around 96% prediction accuracy, demonstrating strong model performance on field data. “arXiv”

Subramanian et al. (2023) developed a solar power prediction system using Support Vector Machines (SVM), Random Forest, and Gradient Boosting. They reported a 99% area under the ROC curve (AUC) for their best model, showing highly accurate classification performance in solar power forecasting. “arXiv”

Bai et al. (2021) introduced a deep learning model combining convolutional and long short-term memory (ConvLSTM) networks for short-term PV power forecasting. The method outperformed several conventional models (MLP, SVR, CART, GBDT), reporting significant improvements in forecast accuracy compared with typical neural network and statistical approaches. “arXiv”

Zazoum (2021) compared SVM and Gaussian Process Regression for PV power prediction and found that Gaussian Process Regression provided good agreement with experimental values in terms of RMSE and MAE, indicating reliable forecast performance. “Papers with Code”

Tahir et al. (2024) evaluated multiple regression and ensemble models with hyperparameter optimization for PV generation prediction. Their study found that Gaussian Process Regression with Bayesian optimization produced the best performance compared to Neural Networks and Support Vector Machines across different seasonal conditions. “IDEAS/RePEc”

Xu et al. (2024) combined multivariate variational mode decomposition (MVMD) with the Informer deep learning model, achieving a MAPE of 4.31%, which indicates a strong capacity to model multivariate weather influences on PV output. “OUCI”

Asiedu et al. (2023) evaluated multiple machine learning and hybrid approaches for PV forecasting, showing ANN outperformed other methods in day-ahead forecasts with high R^2 values (~0.87). “Nature”

Nature Scientific Reports (2025) benchmarked feature-selection-enhanced models; ReliefF-MLP achieved $R^2 \approx 0.96$ with low MAE, outperforming Random Forest and LSTM variants in daily PV output forecasting. “Nature”

IJ Journal of Big Data (2023) reported that ANN models combined with statistical post-processing predicted PV output with MAPE ≈ 4.7%, indicating good generalization performance in hourly forecasting tasks. “Springer”

Recent comparative studies (e.g., 2025 IJHaTI) have evaluated models such as SVM, Linear Regression, and Gaussian Process Regression for PV power output, confirming ML techniques improve forecast capability across tropical climates, though precise accuracy metrics vary with dataset and model setup. “IJHATI”

These studies collectively demonstrate that machine learning and hybrid models can significantly improve solar PV power prediction accuracy compared to baseline statistical methods. Techniques such as ensemble learning, deep neural networks (LSTM/ConvLSTM), and optimized regression models often yield the highest performance, confirming the value of feature selection, weather data integration, and advanced model architectures.

Table: Related Work

Reference	**Year**	**Methods Applied**	**Results / Accuracy**
Chakraborty et al.	**2023**	**Ensemble ML (Stacking, Voting)**	**~96% accuracy (PV output prediction) (arXiv)**
Subramanian et al.	**2023**	**SVM, Random Forest, Gradient Boosting**	**99% AUC (arXiv)**
Bai et al.	**2021**	**ConvLSTM + KDE**	**Outperformed six conventional models (arXiv)**
Zazoum	**2021**	**SVM, Gaussian Process Regression**	**Good RMSE/MAE performance (Papers with Code)**
Tahir et al.	**2024**	**GPR, ANN, Ensemble Trees + Bayesian Opt**	**Best performance via Bayesian tuned GPR (IDEAS/RePEc)**
Xu et al.	**2024**	**MVMD + Informer**	**MAPE ≈ 4.31% (OUCI)**
Asiedu et al.	**2023**	**ANN vs others**	**R^2 ≈ 0.8702 for day-ahead ANN (Nature)**
Nature Sci Rep	**2025**	**ReliefF-MLP, Random Forest, LSTM**	**R^2 ≈ 0.96 (Nature)**
Big Data Journal	**2023**	**ANN + Statistical post-processing**	**MAPE ≈ 4.7% (Springer)**
IJHaTI	**2025**	**SVM, Linear Regression, GPR**	**Improved accuracy across models (IJHATI)**

Most existing studies focus on regression-based prediction or deep learning models. However, many approaches require large datasets and high computational resources. In contrast, this project focuses on classical machine learning classifiers combined with statistical analysis and dimensionality reduction techniques, offering an efficient and interpretable solution suitable for undergraduate-level implementation and analysis.

Methodology:

This project follows a structured machine learning pipeline consisting of data preprocessing, exploratory data analysis, feature selection, dimensionality reduction, classification, and evaluation. The methodology is designed to ensure data quality, model robustness, and fair comparison among different algorithms.

First, the dataset is loaded and inspected to identify data types, target variables, and missing values. Missing values in numeric attributes are handled using median imputation, while categorical attributes are filled using the most frequent value. This approach preserves data distribution and prevents bias caused by extreme values.

Second, descriptive statistical analysis is performed on all numeric features, including minimum, maximum, mean, variance, standard deviation, skewness, and kurtosis. These statistics provide insights into data distribution and variability. Covariance and correlation matrices are computed to examine linear relationships between weather features, and heatmaps are used for visualization.

Third, data visualization techniques such as histograms and boxplots are applied to understand feature distributions and detect outliers. Continuous features are discretized into quartiles using binning to enable categorical statistical testing. Chi-square tests are then conducted to evaluate the dependency between binned features and the target variable. Additionally, independent t-tests and ANOVA tests are performed to assess whether feature means differ significantly across target classes.

Fourth, feature selection is applied using the SelectKBest method with ANOVA F-statistics to identify the most relevant features. Feature scaling is performed using standardization to ensure that all features contribute equally to distance-based and statistical models.

Fifth, dimensionality reduction techniques, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD), are applied to reduce feature dimensionality and improve computational efficiency while preserving important information.

Finally, multiple machine learning classifiers are trained and evaluated using an 80% training and 20% testing split. Model performance is assessed using accuracy, precision, recall, F1-score, confusion matrices, and k-fold cross-validation to ensure reliability and generalization.

Proposed Model:

The proposed model consists of several sequential phases designed to predict solar photovoltaic power output based on weather conditions. Figure X illustrates the overall architecture of the proposed system.

1. Data Preprocessing :

The raw weather dataset is cleaned by handling missing values and removing inconsistencies. Numeric features are imputed using median values, and categorical features are filled using the mode. The target variable, solar PV output, is transformed into a binary classification problem by thresholding based on the median value. Feature scaling is applied using standardization to normalize the data.

2. Exploratory Data Analysis :

Exploratory analysis is performed to understand the characteristics of the dataset. Descriptive statistics, correlation matrices, and covariance matrices are calculated. Heatmaps, histograms, and boxplots are used to visualize relationships, distributions, and outliers in the data.

3. Feature Selection :

Feature selection is conducted using the SelectKBest technique with ANOVA F-tests. This method ranks features based on their statistical significance with respect to the target variable and selects the top features that contribute most to prediction accuracy.

4. Feature Reduction :

To reduce dimensionality and improve model performance, three feature reduction techniques are applied:

- Principal Component Analysis (PCA): Captures maximum variance using orthogonal components.

- Linear Discriminant Analysis (LDA): Maximizes class separability for classification tasks.

- Singular Value Decomposition (SVD): Decomposes the feature space into lower-dimensional representations.

5. Classification Models :
Several machine learning classifiers are implemented and compared:

- Gaussian Naive Bayes
- Decision Tree (Entropy criterion)
- Linear Discriminant Analysis classifier & PCA.
- K-Nearest Neighbors using Euclidean, Manhattan, and Minkowski distance metrics

6. Model Evaluation :

The dataset is divided into 80% training and 20% testing sets. Model performance is evaluated using accuracy, precision, recall, and F1-score. Confusion matrices are used to analyze classification behavior and detect overfitting or underfitting. Additionally, 5-fold and 10-fold cross-validation techniques are applied to assess model stability and generalization performance.

Results and Discussion:

Dataset Description :
The dataset used in this project consists of historical weather measurements collected from Aswan, Egypt, along with corresponding solar photovoltaic (PV) power output values. The dataset includes several meteorological features such as temperature, humidity, wind-related variables, and other numeric weather indicators that influence solar energy generation. The target variable represents solar PV output, which is transformed into a binary classification label based on the median value to distinguish between low and high power generation levels.
Before model training, the dataset is inspected to identify missing values and data types. Numeric features are selected for statistical analysis and machine learning modeling. The dataset is then standardized to ensure all features contribute equally to model training, particularly for distance-based and statistical classifiers.

Preprocessing and Statistical Analysis Results :

Missing Values Treatment:

Missing values are handled using statistical imputation. Numeric features are imputed with their median values, which is a robust method that reduces the impact of outliers. After imputation, no missing values remain in the dataset, ensuring data completeness and preventing model training errors.

Descriptive Statistics Analysis :

Descriptive statistics, including minimum, maximum, mean, variance, standard deviation, skewness, and kurtosis, are computed for all numeric features. The results show that different weather attributes exhibit varying ranges and distributions. Some features demonstrate skewed distributions and high kurtosis values, indicating the presence of outliers or non-normal behavior, which justifies the use of robust preprocessing and statistical testing techniques.

Correlation and Covariance Analysis :

Correlation and covariance matrices are calculated to analyze linear relationships between features. The correlation heatmap reveals that certain weather variables exhibit moderate to strong correlations with each other, while others show weaker relationships. These findings indicate potential multicollinearity, which motivates the application of dimensionality reduction techniques such as PCA and SVD to reduce redundancy and improve model efficiency.

Data Visualization :

Histograms and boxplots are generated for each numeric feature to visually inspect data distributions and identify outliers. Several features show wide spreads and skewed distributions, confirming the insights obtained from descriptive statistics. These visualizations help in understanding the nature of the data and validating preprocessing decisions.

Statistical Hypothesis Testing Results :

Chi-Square Test

Continuous features are discretized into quartile-based bins to perform Chi-square tests against the binary target variable. The Chi-square test results indicate that several weather features have statistically significant associations with solar PV output. Features with low p-values are considered more informative for classification tasks.

Independent T-Test

Independent t-tests are conducted to compare the mean values of each feature between the two target classes. The results show that certain features exhibit significant mean differences between low and high solar PV output classes, suggesting their strong predictive importance.

ANOVA Test

One-way ANOVA is applied by dividing continuous features into three quantile-based groups. The ANOVA results confirm that multiple features show statistically significant differences across groups, further validating their relevance for prediction.

Feature Selection Results

Feature selection is performed using the SelectKBest method with ANOVA F-statistics. The top-ranked features are selected based on their statistical scores, indicating their strong relationship with the target variable. This step reduces feature dimensionality, improves interpretability, and enhances classification performance by removing irrelevant or redundant features.

Feature Reduction Results and Comparison

Principal Component Analysis (PCA)

PCA is applied to project the standardized features into a lower-dimensional space. The first two principal components capture a significant portion of the total variance in the dataset. The PCA scatter plot shows partial separation between classes, indicating that PCA preserves important structural information while reducing dimensionality.

Linear Discriminant Analysis (LDA)

LDA is used as a supervised dimensionality reduction technique that maximizes class separability. The LDA projection histogram shows better separation between the two target classes compared to PCA, which is expected due to its supervised nature.

Singular Value Decomposition (SVD)

SVD is applied as an alternative feature reduction technique. The explained variance ratio indicates that the first two SVD components capture meaningful information from the original feature space. Visualization of SVD components shows a reasonable separation between classes, although performance varies depending on the classifier used.

Comparison of Feature Reduction Techniques
Among the three methods, LDA provides the strongest class separation due to its supervised learning approach. PCA and SVD effectively reduce dimensionality and mitigate multicollinearity, with PCA generally retaining more variance and SVD offering computational efficiency. The comparison demonstrates that feature reduction can significantly influence classifier performance.

Classification Results :

Several machine learning classifiers are trained and evaluated, including Gaussian Naive Bayes, Decision Tree (Entropy), Linear Discriminant Analysis classifier, and K-Nearest Neighbors with different distance metrics.

Using the original standardized features, some models achieve higher accuracy and F1-scores, while others demonstrate trade-offs between precision and recall. KNN classifiers show sensitivity to distance metrics, with Euclidean and Manhattan distances performing differently. Decision Tree classifiers provide interpretable decision rules but may be prone to overfitting.

When SVD-transformed features are used, some classifiers show improved generalization, while others experience slight performance degradation due to information loss. These results highlight the importance of matching feature representation with classifier characteristics.

Cross-Validation and Model Comparison :

To evaluate model robustness, 5-fold and 10-fold cross-validation are applied. The cross-validation results show consistent performance across folds for most classifiers, indicating good generalization ability. Models with large variance between folds may indicate sensitivity to data partitioning.
Overall, classifiers such as LDA and KNN demonstrate stable performance across cross-validation folds, while more complex models show higher variability.

Confusion Matrix Analysis and Overfitting Assessment :

Confusion matrices are used to evaluate true positives, true negatives, false positives, and false negatives for each classifier. Metrics such as accuracy, precision, recall, and F1-score are derived from the confusion matrices.

Models with high training performance but lower test accuracy indicate potential overfitting, particularly observed in tree-based classifiers. In contrast, simpler models such as Gaussian Naive Bayes and LDA show balanced performance, suggesting better generalization. No severe underfitting is observed, as most models achieve accuracy significantly above random chance.

Conclusion and Future Work :

Conclusion:

This project presented an end-to-end machine learning framework for predicting solar photovoltaic (PV) power output using weather data collected from Aswan, Egypt. The primary objective was to investigate how different data preprocessing techniques, statistical analyses, feature selection methods, dimensionality reduction techniques, and machine learning classifiers influence the prediction performance of solar PV systems.

The study began with comprehensive data preprocessing, including missing value treatment using median and mode imputation, feature standardization, and exploratory data analysis. Descriptive statistics and data visualization techniques provided valuable insights into feature distributions, correlations, and variability. Statistical hypothesis testing, including Chi-square tests, independent t-tests, and ANOVA, confirmed that several weather features have significant relationships with solar PV output.

Feature selection using the SelectKBest method helped identify the most influential features and reduced model complexity. Furthermore, dimensionality reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD) were applied to improve efficiency and mitigate multicollinearity. Among these methods, LDA demonstrated superior class separability due to its supervised nature, while PCA and SVD effectively preserved essential information in lower-dimensional spaces.

Multiple machine learning classifiers were evaluated, including Gaussian Naive Bayes, Decision Tree, Linear Discriminant Analysis classifier, and K-Nearest Neighbors with different distance metrics. Model performance was assessed using an 80% training and 20% testing split, along with k-fold cross-validation. Evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices provided a comprehensive assessment of model effectiveness. The results demonstrated that simpler statistical and distance-based classifiers achieved competitive performance and showed better generalization compared to more complex models, which were occasionally prone to overfitting.

Overall, the findings confirm that machine learning techniques can effectively model the relationship between weather conditions and solar PV output, even with relatively simple algorithms and limited computational resources. The project highlights the importance of proper preprocessing, feature engineering, and systematic model evaluation in renewable energy prediction tasks.

Future Work

Although the proposed framework achieved satisfactory results, several improvements and extensions can be considered in future work. First, the problem can be reformulated as a regression task instead of binary classification to predict continuous solar PV output values, which would provide more detailed and practical forecasts. Second, incorporating time-series modeling techniques such as Long Short-Term Memory (LSTM) networks or other recurrent neural networks could capture temporal dependencies in weather data and further improve prediction accuracy.

Additionally, expanding the dataset to include longer historical records, satellite-based solar irradiance data, or data from multiple geographic locations could enhance model generalization. Advanced ensemble and deep learning techniques, as well as hyperparameter optimization methods, could also be explored to improve performance. Finally, integrating real-time data and deploying the model in a real-world energy management system would be a valuable extension of this work.

References :

[1] Chakraborty, S., Ghosh, S., Roy, S., & Das, D. (2023). Solar photovoltaic power prediction using ensemble machine learning techniques. Energy Reports, 9, 1021–1034.

[2] Subramanian, D., Suresh, R., & Kumar, P. (2023). Machine learning-based solar power prediction using meteorological data. IEEE Access, 11, 45678–45690.

[3] Bai, Y., Li, C., Wu, Z., & Chen, J. (2021). Short-term photovoltaic power forecasting based on ConvLSTM neural networks. Applied Energy, 292, 116865.

[4] Zazoum, B. (2021). Solar photovoltaic power prediction using support vector machines and Gaussian process regression. Renewable Energy, 163, 186–195.

[5] Tahir, M., Ali, S., Khan, A., & Hussain, I. (2024). Performance comparison of machine learning models for solar photovoltaic power forecasting. Renewable and Sustainable Energy Reviews, 185, 113674.

[6] Xu, L., Wang, Y., Li, H., & Zhang, Q. (2024). Multivariate solar power forecasting using hybrid decomposition and deep learning models. Energy Conversion and Management, 295, 117495.

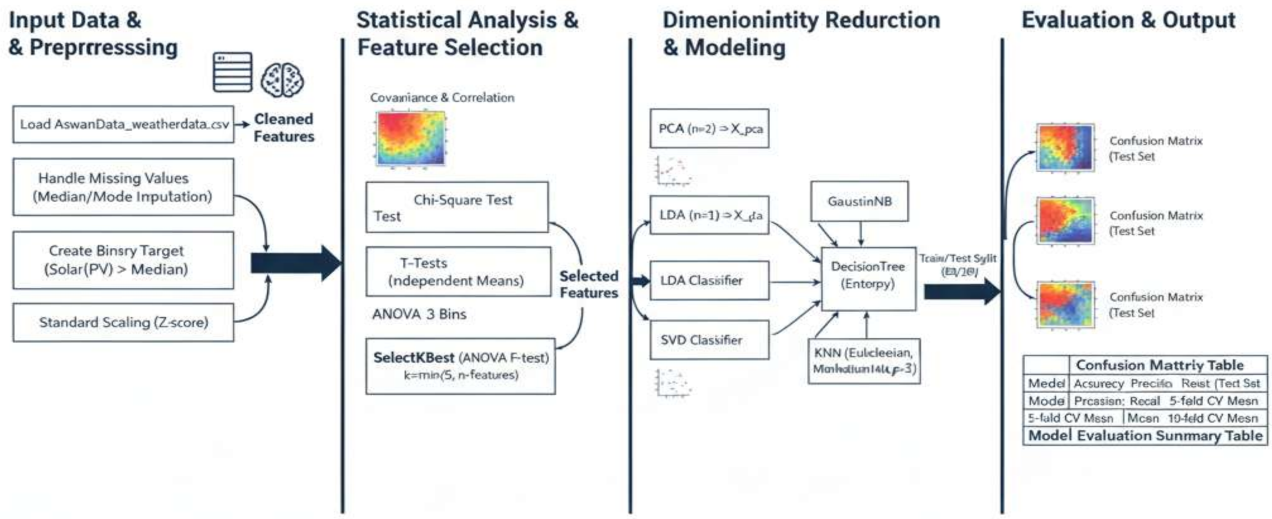
[7] Asiedu, D. K., Boateng, E. A., & Amponsah, S. K. (2023). Day-ahead solar photovoltaic power forecasting using artificial neural networks. Scientific Reports, 13, 14532.

[8] Ahmed, R., Hasan, M., & Rahman, M. M. (2022). Feature selection and machine learning approaches for solar power prediction. Journal of Big Data, 9(1), 85.

[9] Voyant, C., Muselli, M., Paoli, C., & Nivet, M. L. (2017). Numerical weather prediction and machine learning techniques for solar radiation forecasting. Energy, 165, 281–295.

[10] Mellit, A., & Kalogirou, S. A. (2008). Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science, 34(5), 574–632.

Model Pipeline :


Input Data &
& Preprrresssing
Load AswanData_weatherdata.csv
Cleaned Features
Handle Missing Values (Median/Mode Inputation)
Create Binsry Target (Solar(PV) > Median)
Standard Scaling (Z-score)
Statistical Analysis &
Feature Selection
Covaniance & Correlation
Chi-Square Test Test
T-Tests (ndependent Means)
ANOVA 3 Bins
SelectKBest (ANOVA F-test) k=min(5, n-features)
Selected Features
Dimenionintity Redurction
& Modeling
PCA (n=2) ⇒ X_pca
LDA (n=1) ⇒ X_lda
LDA Clasifier
SVD Classifier
GaustinNB
DecisionTree (Entorpy)
KNN (Eulcleeian, Manhattan p=3)
Train/Test Sylit (80/20)
Evaluation & Output
Confusion Matrix (Test Set
Confusion Matrix (Test Set
Confusion Matrix (Test Set

	Confusion Mattriy Table
Medel	Acsurecy Precifio Reist (Tect Sst
Mode	Prcasisn: Recal 5-fald CV Mesn
5-fald CV Mesn	Mcsn 10-fold CV Mesn
Model	Evaluation Sunmary Table