Description of the dataset:

The project aims to use machine learning classifiers to categorize defects in photovoltaic (PV) systems. Moreover, the dataset includes operational data from a grid-tie solar power plant that uses a variety of factors, including voltages, currents, radiation levels, and temperature, to differentiate between normal and malfunctioning operations. Creating and assessing classification models for fault detection is the main goal. Also, the main characteristics of the dataset are the voltage between two cell strings, the currents produced by the two strings, the solar irradiation, and the temperature of the PV module. The fault label is divided into five classes: 0 for normal operation, 1 for short circuit, 2 for degradation, 3 for open circuit, and 4 for shadowing.

Numerical summaries of the PV data:**A screenshot of a computer

AI-generated content may be incorrect.**

Using the describe() function in Python, it provides a comprehensive overview of key features in the dataset related to fault detection in a photovoltaic (PV) system. Furthermore, the dataset has 25,000 observations with seven variables, including voltages across two strings of cells (String 1 and String 2), currents from both strings, solar irradiance, PV module temperature, and the fault label. Additionally, notable observations include a variety of voltage and current values,  in string 1 with mean voltages around 187.74V and 5.13A. The irradiance shows a mean of approximately 661.53 W/m², however, an unusual minimum value of -0.0381 W/m² raises questions about potential errors or outliers in the data. The fault label variable represents five fault classes with the majority centered around the 'Degradation' category.

Graphical summaries of the PV data:

A visual representation of the distribution of irradiance values across different fault categories in the Photovoltaic (PV) dataset. The x-axis represents the fault labels ranging from 0 to 4, each corresponding to a specific fault type. The y-axis showcases the values of irradiance, with each box representing the quartile range of irradiance values for a given fault label. The central line within each box denotes the median irradiance, while the whiskers extend to highlight the range of values. This graph allows for a quick comparison of irradiance distributions among different fault categories, aiding in identifying potential patterns that may contribute to fault detection.

A chart with multiple colored boxes

AI-generated content may be incorrect.

The histograms show the distribution of currents in Strings 1 and 2. The average number of data within a given range is shown by each bar. These histograms for currents display the amount at which various current values appear in each string. The histograms indicate that String 1's current values are more evenly distributed, while String 2's distribution is somewhat positively skewed.

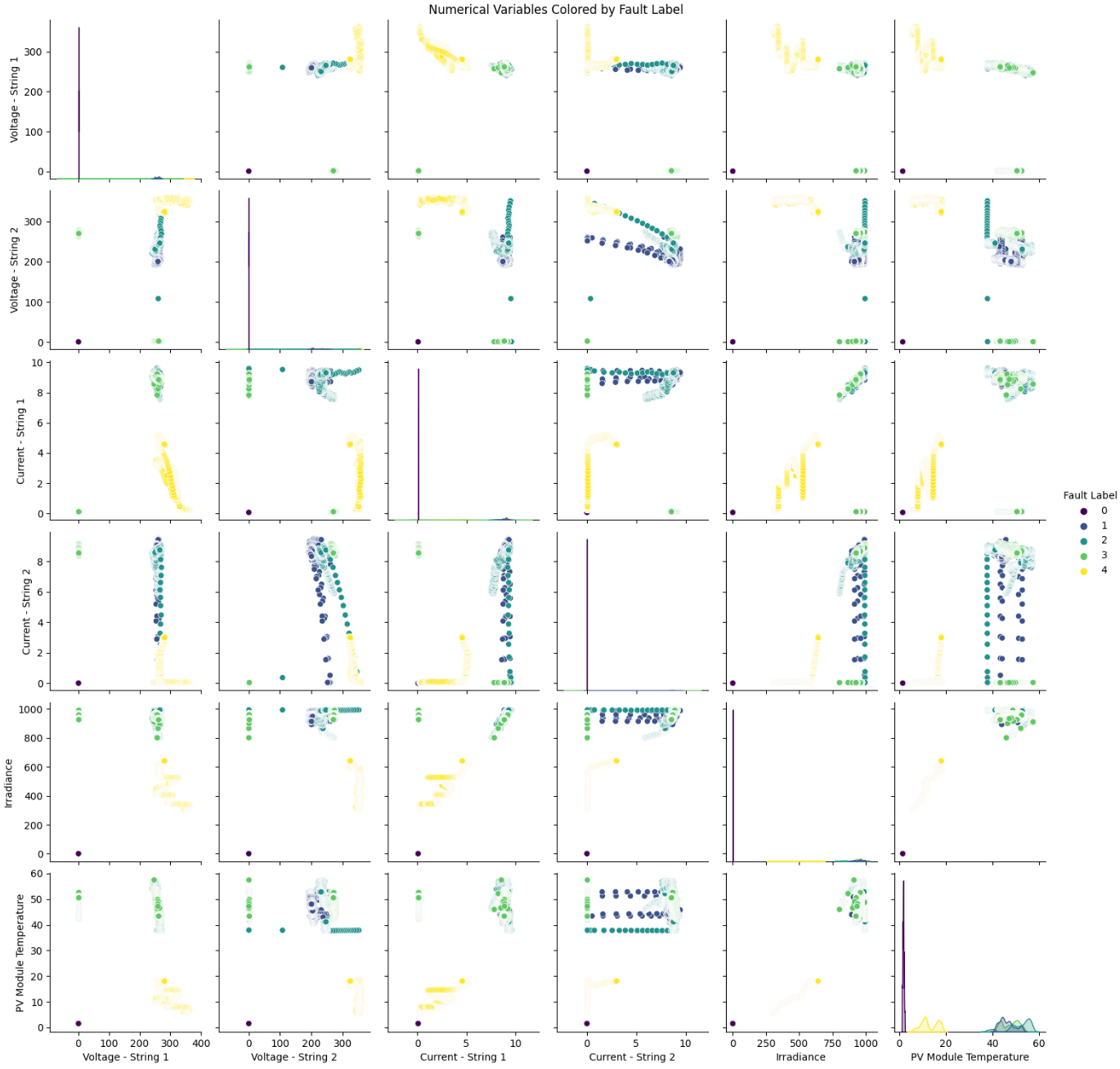
A graph of current and current

AI-generated content may be incorrect.

The correlation coefficients between each pair of features in the dataset are displayed visually in the heat map. The degree and direction of a linear link between two variables are measured by correlation. The degree of correlation is shown by the color intensity in the heat map; warmer colors (reds) denote a positive correlation, while cooler colors (blues) denote a negative correlation. This plot allows the identification of feature correlations.  In this case, the heat map indicates that the characteristics often have low correlation, which is advantageous for creating strong classification models.

A screenshot of a graph

AI-generated content may be incorrect.

The pair plot displays scatterplots for pairs of numerical variables and diagonal histograms for each variable. Also, points are color-coded based on the fault label, allowing us to observe relationships between variables and their distribution within each fault category.

The scatter plots offer a detailed examination of relationships between key variables in the fault detection dataset for a Photovoltaic (PV) system. Additionally, the scatter plot visualizes the correlation between voltages across String 1 and String 2, utilizing color to distinguish fault labels. This aids in identifying any discernible patterns or clusters associated with different fault categories. The color-coded fault labels enhance the interpretation, revealing insights into the role of these variables in fault classification.

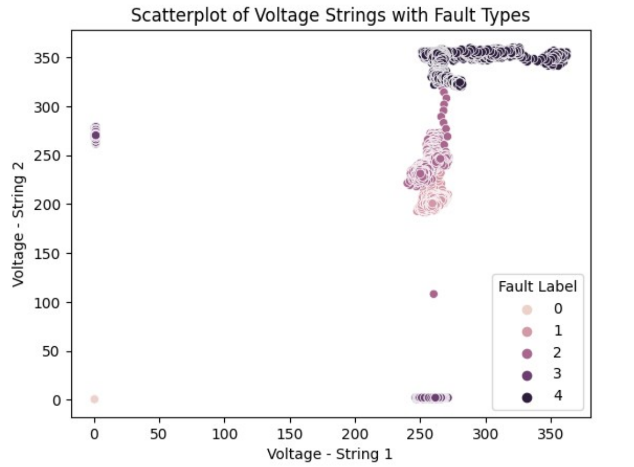


Table of classification results:

A thorough analysis of the effectiveness of several classifiers in defect detection inside a photovoltaic system is given in the classification results table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test Split 1 | | Test Split 2 | | Test Split 3 | | Average | |
| Classifier | Norm 1 | Norm 2 | Norm 1 | Norm 2 | Norm 1 | Norm 2 | Norm 1 | Norm 2 |
| k-NN | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9993 | 0.9993 | 2.9984 | 2.9984 |
| Decision Trees | 0.9996 | 0.9996 | 0.9993 | 0.9993 | 0.9995 | 0.9995 | 2.9984 | 2.9984 |
| Naïve Bayes | 0.9972 | 0.9972 | 0.9957 | 0.9957 | 0.9961 | 0.9961 | 2.989 | 2.989 |
| SVM (polyn.) | 0.9991 | 0.9993 | 0.9981 | 0.9983 | 0.9984 | 0.9987 | 2.9956 | 2.9963 |
| SVM (RBF) | 0.9991 | 0.9991 | 0.9981 | 0.9981 | 0.9987 | 0.9987 | 2.9959 | 2.9959 |
| Neural Networks | 0.9995 | 0.9992 | 0.9992 | 0.9989 | 0.9992 | 0.9992 | 2.9979 | 2.9973 |

Notably, over three distinct test splits and two normalization procedures (Norm 1 and Norm 2), k-NN, Decision Trees, SVM with Polynomial Kernel, SVM with RBF Kernel, and Neural Networks consistently display high accuracy, ranging from 99.57% to 99.96%. While SVMs and Neural Networks continue to achieve comparable levels of accuracy, the k-NN and Decision Trees classifiers demonstrate strong and steady performance. Naïve Bayes exhibits its applicability for fault classification despite achieving a somewhat lower accuracy than some other classifiers. Together, these findings demonstrate how well the selected classifiers performed at correctly identifying fault labels in the photovoltaic system dataset, offering a solid basis for fault detection and classification tasks.