

Project Summaries

Weather Data Analysis and Income Prediction

This document contains summaries of two separate data analysis projects. The first project focuses on predicting temperature, and the second focuses on predicting income level.

1. Weather Data Analysis and Temperature Prediction Project Summary

Objective:

The objective of this project is to analyze historical weather data and develop regression models to predict future temperatures. Specifically, the models are designed to use today's weather data to predict tomorrow's temperature.

Methodology:

The project involves several key steps:

1. **Data Cleaning:** This involves handling any missing or inconsistent data in the historical weather dataset.
2. **Feature Engineering:** This involves creating new relevant features from the existing data, such as calculating daily averages of weather variables.
3. **Model Selection:** This involves evaluating linear and polynomial regression models for temperature prediction.
4. **Model Evaluation:** This involves assessing the performance of the models using metrics such as Mean Squared Error (MSE) and R-squared (R^2).

Results:

The multiple linear regression model demonstrated the best performance, achieving an MSE of 4.22 and an R^2 of 0.945. This indicates that the model explains a large proportion of the variance in the temperature data and has a relatively low prediction error. Polynomial regression models did not outperform the linear regression models.

Conclusion:

This project analyzed historical weather data to train regression models for predicting future temperatures. The multiple linear regression model performed best, demonstrating the effectiveness of using statistical models to forecast future temperatures based on previous weather conditions.

2. Income Level Prediction with Logistic Regression

Objective:

This project aims to analyze demographic data to predict whether an individual's income exceeds \$50,000. The project focuses on applying logistic regression for this classification task.

Methodology:

The project follows these steps:

1. **Exploratory Data Analysis (EDA):** The dataset is loaded, cleaned (handling missing values, data types), and visualized to understand the data distribution and relationships between variables.
2. **Preprocessing:** Categorical variables are encoded (using one-hot and ordinal encoding), data is cleaned, features are scaled, and the dataset is split into training and testing sets.
3. **Modeling with Logistic Regression:** A logistic regression model is trained on the preprocessed training data.
4. **Model Evaluation:** The model's performance is evaluated using metrics such as accuracy, F1-score, precision, recall, and ROC AUC. A confusion matrix is also generated and visualized.

Results:

The logistic regression model achieved an accuracy of 82.91% on the test set. Additional metrics include a precision of 0.66, recall of 0.67, and an F1-score of 0.66. These metrics indicate that the model performs reasonably well in predicting income level. However, the dataset exhibits a significant class imbalance.

Conclusion:

This project used logistic regression to predict income levels from demographic data. The model shows reasonable predictive power, but performance is affected by class imbalance. Addressing this imbalance may improve the results.