**Thesis Report on Bike Sharing Dataset Analysis**

**Introduction**

This report details the comprehensive analysis conducted on the Bike Sharing Dataset, with the aim of predicting and understanding bike rental patterns. The dataset includes various features such as temperature, humidity, windspeed, and count of total bike rentals. We applied multiple machine learning models and statistical tests to explore the relationships within the data, predict rental counts, and assess the performance of different models.

**Objectives**

The primary objectives of this analysis were:

1. To predict the total number of bike rentals using regression techniques.
2. To classify rental counts as high or low using various classification models.
3. To apply statistical tests to understand differences in rental patterns across different conditions.

**Data Preparation**

The dataset was preprocessed to extract relevant features:

* **Independent Variables**: Temperature, humidity, and windspeed.
* **Dependent Variables**:
  + For regression: The total count of bike rentals (cnt).
  + For classification: A binary variable indicating high or low rentals (high\_rentals), based on whether the rental count was above or below the median.

**Modeling and Analysis**

1. **Linear Regression**
   * **Objective**: Predict the total number of bike rentals.
   * **Approach**: A linear regression model was trained using features such as temperature, humidity, and windspeed to predict cnt.
   * **Outcome**: The model provided a continuous prediction of rental counts. The performance was evaluated using metrics like Mean Squared Error (MSE) and R^2 Score.
2. **Logistic Regression**
   * **Objective**: Classify rental counts as high or low.
   * **Approach**: A logistic regression model was trained to predict the binary high\_rentals variable using the same features.
   * **Outcome**: The model's performance was evaluated using a confusion matrix and classification report, indicating accuracy, precision, recall, and F1-score.
3. **Support Vector Machine (SVM)**
   * **Objective**: Classify rental counts as high or low.
   * **Approach**: An SVM with a linear kernel was applied to classify rentals, focusing on maximizing the margin between the two classes.
   * **Outcome**: The SVM model's performance was compared with Logistic Regression, using similar evaluation metrics.
4. **Random Forest**
   * **Objective**: Classify rental counts as high or low.
   * **Approach**: A Random Forest classifier was trained, which creates an ensemble of decision trees to improve prediction accuracy and reduce overfitting.
   * **Outcome**: The model was evaluated with the same metrics and showed better generalization compared to individual trees.
5. **Decision Tree**
   * **Objective**: Classify rental counts as high or low.
   * **Approach**: A single decision tree model was trained to classify rentals based on feature splits.
   * **Outcome**: The decision tree provided an interpretable model with decent accuracy but was prone to overfitting.

**Statistical Tests**

1. **T-Test**
   * **Objective**: Compare mean bike rentals on weekdays vs. weekends.
   * **Approach**: An independent T-test was conducted to check if the difference in mean rentals between weekdays and weekends is statistically significant.
   * **Outcome**: The results showed whether or not there was a significant difference, helping to understand usage patterns based on the day of the week.
2. **McNemar's Test**
   * **Objective**: Compare the performance of two classification models (e.g., Logistic Regression vs. SVM).
   * **Approach**: McNemar’s test was applied on the contingency table of predictions from the two models.
   * **Outcome**: The test determined if there was a significant difference in the models' performance on the same data, helping to choose the better model.
3. **F-Test (ANOVA)**
   * **Objective**: Compare variance in bike rentals across different seasons.
   * **Approach**: ANOVA was used to compare the mean rental counts across different seasons to determine if any seasonal effect was significant.
   * **Outcome**: The test highlighted whether seasonality had a significant impact on bike rentals, which could be crucial for planning and resource allocation.

**Results and Discussion**

* **Regression Analysis**: The linear regression model provided a decent fit for predicting total bike rentals, with reasonable accuracy. However, the model’s performance could be affected by the linearity assumption.
* **Classification Models**: The Random Forest classifier outperformed other models (Logistic Regression, SVM, Decision Tree) in terms of accuracy and robustness. SVM provided a strong alternative with clear decision boundaries, while Decision Tree, though interpretable, was prone to overfitting.
* **Statistical Tests**:
  + The T-test revealed significant differences in rentals between weekdays and weekends, suggesting different usage patterns.
  + McNemar’s test confirmed that there was a statistically significant difference between the Logistic Regression and SVM models' performance.
  + The F-test demonstrated that seasonality significantly affects bike rentals, with variations in demand across different seasons.

**Conclusion**

This analysis provided valuable insights into bike rental patterns using various machine learning models and statistical tests. The findings suggest that Random Forest is the most effective model for classification tasks in this context, while seasonality and day-of-week effects are significant factors in rental behavior. These results can inform better decision-making for managing bike-sharing systems, including resource allocation and planning for peak times.

The report demonstrates the application of predictive modeling and statistical testing, contributing to the broader understanding of urban mobility patterns, which is crucial for smart city planning and sustainable transportation solutions.

The analysis highlights the importance of weather and temporal factors in predicting bike rentals. The models and tests used provided valuable insights for understanding and forecasting bike rental patterns, which can be crucial for optimizing bike-sharing systems and resource allocation.