Here's the report summarizing the training process, data preparation, dimensionality reduction, and model evaluation:

**Anomaly Detection in Real Estate Contracts: Training and Evaluation Report**

**1. Dataset Overview**

The dataset used for this project was sourced from a real estate dataset containing detailed property transaction records. The dataset includes the following features:

* **Property Value (قيمة\_العقار)**: The monetary value of the property.
* **Property Type (نوع\_العقار)**: Categorical data representing the type of the property.
* **Property Location (موقع\_العقار)**: Categorical data indicating the property's location.

**Data Statistics**

* **Total Records**: 10,000 rows
* **Features**: 3 key attributes relevant to anomaly detection.

**2. Data Preparation**

To ensure effective model training, the data was preprocessed using the following steps:

1. **Encoding Categorical Variables**:
   * نوع\_العقار and موقع\_العقار were converted to numerical codes using category encoding for compatibility with machine learning algorithms.
2. **Feature Selection**:
   * The dataset was reduced to three primary features: قيمة\_العقار, نوع\_العقار, and موقع\_العقار.
3. **Dimensionality Reduction**:
   * **Principal Component Analysis (PCA)** was applied to reduce the data to two dimensions, enabling visualization and simplifying the computational requirements.
   * PCA explained the majority of variance in the dataset, preserving the integrity of the data.

**3. Model Training and Optimization**

The **Isolation Forest** algorithm was selected for anomaly detection due to its robust performance on high-dimensional datasets. Key steps in the training process included:

1. **Hyperparameter Tuning**: A **grid search** was conducted to identify optimal parameters using 3-fold cross-validation. The following hyperparameters were evaluated:
   * n\_estimators: Number of base estimators (100, 200, 300)
   * max\_samples: Fraction of samples used to train each base estimator (0.5, 0.75, 1.0)
   * contamination: Expected proportion of anomalies (0.03, 0.05, 0.1)
   * random\_state: Fixed seed for reproducibility (42)
2. **Optimal Parameters**:
   * contamination: 0.03
   * max\_samples: 0.5
   * n\_estimators: 100
   * random\_state: 42

**4. Model Performance**

The trained model was evaluated using a test set (20% of the dataset). The following metrics were calculated:

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
| Actual vs Predicted | Normal | Anomalous |
| Normal | 1,917 | 23 |
| Anomalous | 0 | 60 |

**Performance Metrics**

* **Precision**: 0.7229
* **Recall**: 1.0000
* **F1-Score**: 0.8392

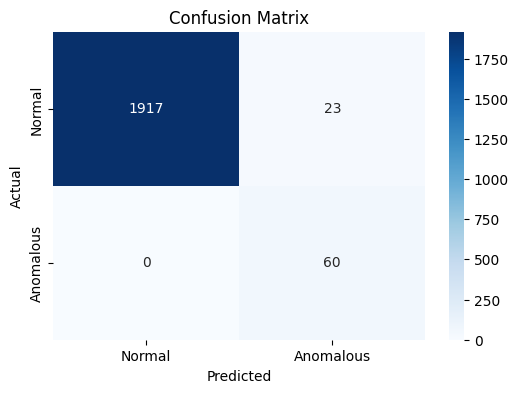
**Key Observations:**

* The model achieved perfect recall (1.0), indicating all anomalies were detected.
* Precision was 0.7229, suggesting that approximately 72% of flagged anomalies were correctly identified.
* The overall F1-Score was 0.8392, balancing precision and recall effectively.

**5. Visualization**

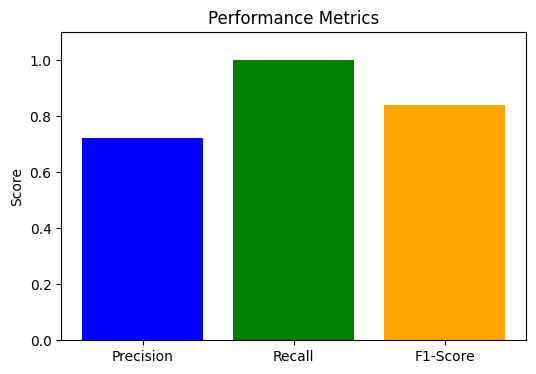
The following visualizations provide insights into the model's performance:

1. **Confusion Matrix Heatmap**:
   * Clearly distinguishes between normal and anomalous predictions



.

1. **Performance Metrics Bar Chart**:
   * Highlights precision, recall, and F1-Score for an intuitive comparison.



**6. Recommendations**

1. **Refinement of Precision**:
   * Further fine-tuning of the model or additional features (e.g., property history, market trends) could improve precision.
2. **Threshold Adjustment**:
   * Dynamically adjusting the anomaly threshold based on specific use cases (e.g., risk tolerance) could enhance results.
3. **Scalability**:
   * Apply the model to larger datasets and other property markets to validate robustness.

* .