Q1 In a rural agricultural region, a group of farmers, supported by a research team, endeavors to understand the correlation between temperature variations and wheat crop yield. The objective is to optimize agricultural practices and enhance wheat production, considering different temperature levels experienced during the growing season.

Context:

- **Location:** The agricultural region experiences varying temperature levels throughout the year, with notable fluctuations during the wheat growing season.
- **Experiment Design:** The research team conducts experiments at five distinct temperature levels to analyze their impact on wheat crop yield.
- **Data Collection:** Measurements are taken for both temperature (x) in degrees Fahrenheit and wheat crop yield (y) in tons per acre.

Experimental Data:

Experiment	Temperature (°F)	Wheat Crop Yield (tons/acre)
1	50	3.3
2	50	2.8
3	50	2.9
4	70	2.3
5	70	2.6
6	70	2.1
7	80	2.5
8	80	2.9
9	80	2.4
10	90	3.0
11	90	3.1
12	90	2.8
13	100	3.3
14	100	3.5
15	100	3.0

Analysis Objective:

Relationship Exploration: Analyze the relationship between temperature variations and heat crop yield to ascertain how temperature fluctuations influence crop production. (2 marks)

Model Development: (3 marks)

1. Choose an appropriate ML algorithm (e.g., Regression models) for prediction based on the selected features.
2. Model the mathematical equation.

Model Evaluation: (3 marks)

1. Train the ML model on the training set and evaluate its performance using loss function.
2. Residual plot analysis to understand model errors.

Conclusion: (3 marks)

1. Summarize the model's performance,
2. Provide insights into the predictive capability and limitations of the ML model.
3. Suggest possible improvements or considerations for future analysis.

Python code (9 marks)

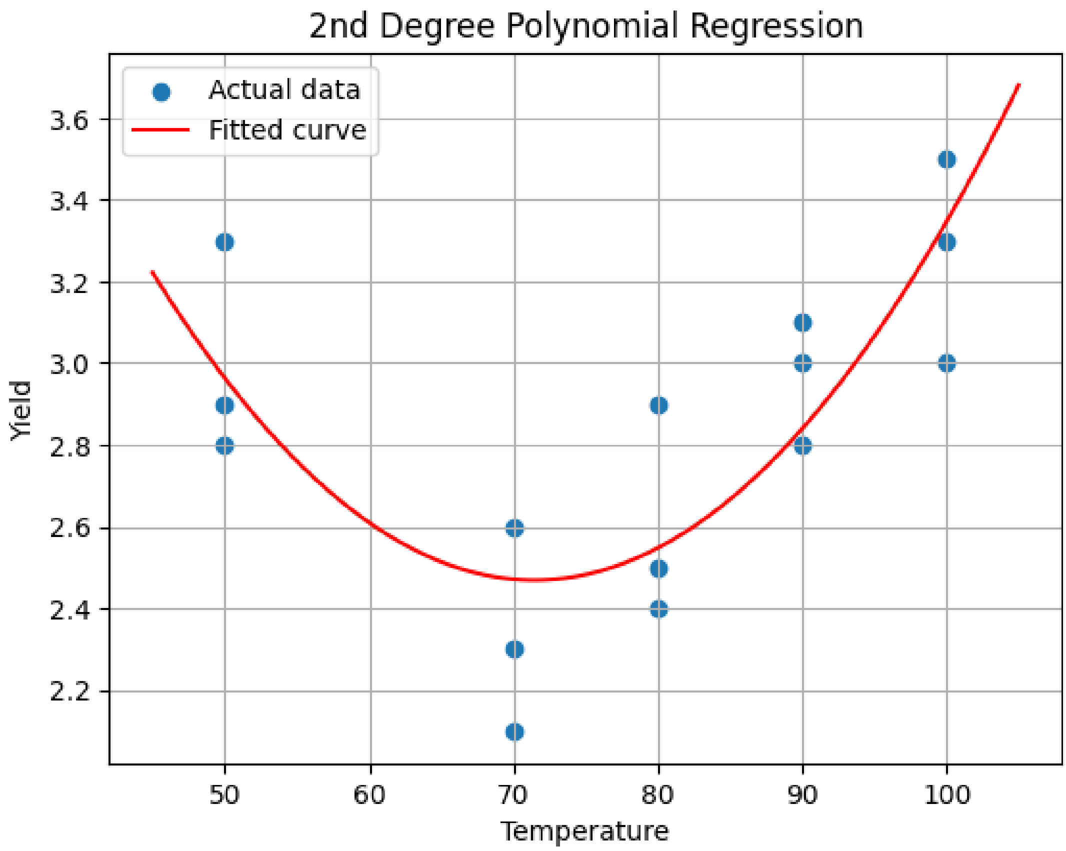
SOLUTION:

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import r2_score, mean_absolute_error,
mean_squared_error
import matplotlib.pyplot as plt

# Data
temperature = np.array([50, 50, 50, 70, 70, 70, 80, 80, 80, 90, 90, 90,
100, 100, 100]).reshape(-1, 1)
yield_data = np.array([3.3, 2.8, 2.9, 2.3, 2.6, 2.1, 2.5, 2.9, 2.4,
3.0, 3.1, 2.8, 3.3, 3.5, 3.0])

# Polynomial regression degree 2
poly = PolynomialFeatures(degree=2)
temperature_poly = poly.fit_transform(temperature)

# Model fitting
lr = LinearRegression()
lr.fit(temperature_poly, yield_data)

# Coefficients of the 2nd degree polynomial equation
coef = lr.coef_
intercept = lr.intercept_
print("Coefficients:", coef)
print("Intercept:", intercept)

# R² score
y_pred = lr.predict(temperature_poly)
```

```
r_squared = r2_score(yield_data, y_pred)
print("R² Score:", r_squared)

# Calculate loss functions
mae = mean_absolute_error(yield_data, y_pred)
mse = mean_squared_error(yield_data, y_pred)
rmse = np.sqrt(mse)
print('MAE:', mae)
print('MSE:', mse)
print('RMSE:', rmse)

# Residual plot
residuals = yield_data - y_pred
plt.scatter(y_pred, residuals)
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.axhline(y=0, color='red', linestyle='--')
plt.show()

# Plotting the fitted curve
plt.scatter(temperature, yield_data, label='Actual data')
plt.xlabel('Temperature')
plt.ylabel('Yield')
plt.title('2nd Degree Polynomial Regression')
plt.grid(True)

# Create points for the fitted curve
x_values = np.linspace(45, 105, 100).reshape(-1, 1)
x_values_poly = poly.transform(x_values)
y_values = lr.predict(x_values_poly)

# Plot the fitted curve
plt.plot(x_values, y_values, color='red', label='Fitted curve')
plt.legend()
plt.show()
```

- **Coefficients and modeled equation**: These are the coefficients of the polynomial equation. For a second-degree polynomial equation of the form $y=ax^2+bx+c$, the coefficients in this case are for x, and the intercept term respectively. In this case:

Coefficients: [0. -0.15371134 0.0010756] Intercept: 7.960481099656514]

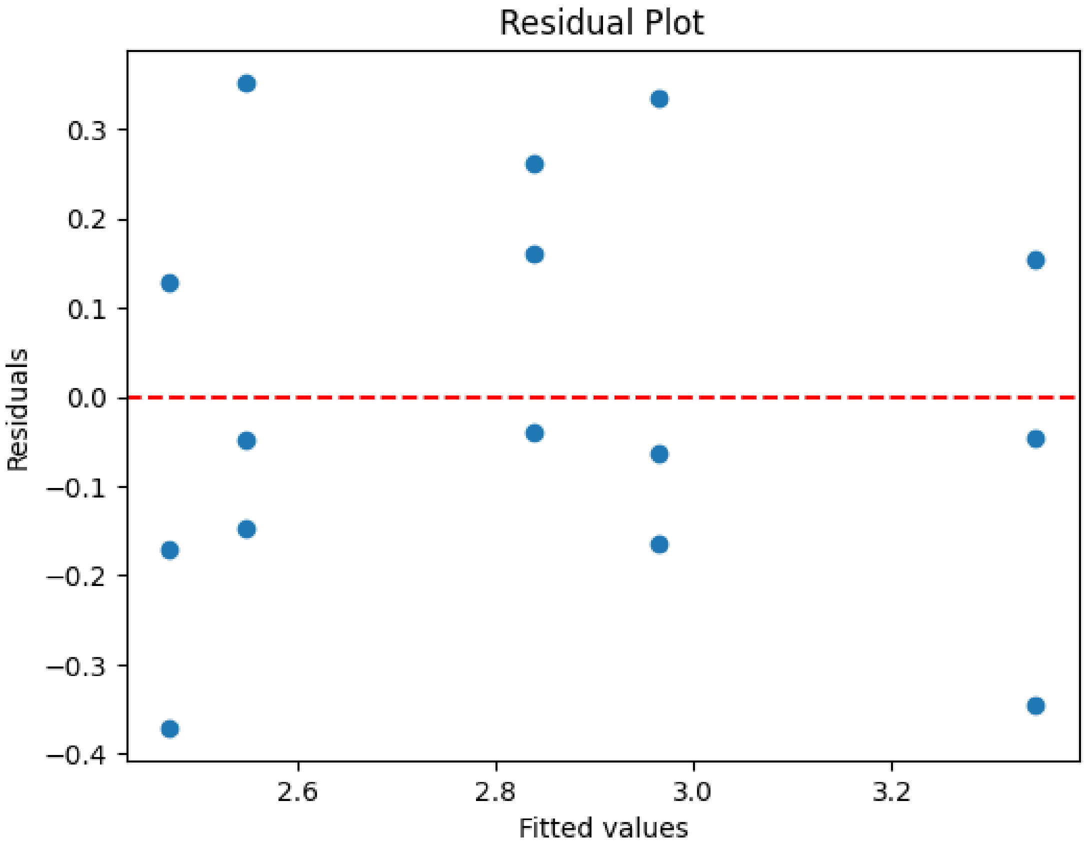


Model Evaluation

- **R² Score (Coefficient of Determination):** It measures the proportion of the variance in the dependent variable (yield) that is predictable from the independent variable (temperature). An R² score of 0.6732 suggests that approximately 67.32% of the variance in yield is explained by the temperature using this model.

- **MAE (Mean Absolute Error):** This measures the average of the absolute errors between predicted and actual values. It indicates the average magnitude of the errors in the predicted values. In this case, it is 0.1859.
- **MSE (Mean Squared Error):** This measures the average of the squared differences between predicted and actual values. It provides an average of the squares of the errors and is useful for understanding the quality of the model's predictions. Here, it's 0.0478.
- **RMSE (Root Mean Squared Error):** It's the square root of the MSE. It represents the standard deviation of the residuals and is interpreted in the same units as the dependent variable. In this case, it's 0.2186.

- Residual plot: A residual plot is a graphical representation used to examine the goodness of fit in regression analysis. It plots the residuals (the differences between observed and predicted values) against the predictor variable(s) used in the model. Residual plot should exhibit a random scatter of points around the horizontal axis, indicating that the model adequately captures the variability in the data. In this case, as data is quite scattered around the horizontal line, so, improved ML model could be used.



Conclusion

These metrics (MAE, MSE, RMSE) provide insights into the accuracy and quality of the regression model, giving an idea of how well the model fits the data. Lower values of MAE, MSE, and RMSE indicate better model performance.

1. **Predictive Capability and Limitations:**

    - The model demonstrates moderate predictive capability by explaining around 67.32% of the yield variance using temperature. However, there might be other factors influencing yield not considered in this model.

    - It's important to note that this model assumes a polynomial relationship between temperature and yield. If the relationship is more complex or nonlinear, a different modeling approach might yield better results.

2. **Possible Improvements and Future Considerations:**

    - Collect more comprehensive data encompassing additional features that could potentially influence yield, such as humidity, precipitation, soil quality, etc.

    - Consider exploring more sophisticated modeling techniques that can capture nonlinear relationships more effectively, such as random forests, gradient boosting, or neural networks.

- Perform feature engineering to extract more meaningful features or transformations of existing features that might better represent the underlying patterns in the data.
- It's crucial to be cautious while selecting the polynomial degree because using a higher degree might lead to overfitting, where the model fits too closely to the training data but performs poorly on new, unseen data.