Predictive Analysis Report on Cervical Cancer Dataset

Objective:

Determine the best combination of features to predict the target variable "Biopsy" in the cervical cancer dataset.

Feature Combinations Evaluated:

1. "Schiller", "Hinselmann", and "Age"
2. "Schiller", "Hinselmann", and "Hormonal Contraceptives (years)"

Analysis & Findings:

1. Features: "Schiller", "Hinselmann", and "Age":
   - Model Performance: Achieved an accuracy of 96.41%.
   - Observations from Visualizations:
      - Both "Schiller" and "Hinselmann" displayed binary outcomes.
      - "Age" represented the age distribution of the individuals in the dataset.
      - Scatter plots showcased a clearer distinction in biopsy outcomes for the "Schiller" feature as compared to "Hinselmann" when set against "Age".

2. Features: "Schiller", "Hinselmann", and "Hormonal Contraceptives (years)":
   - Model Performance: Achieved an accuracy of 94.02%.
   - Observations from Visualizations:
      - "Schiller" and "Hinselmann" again displayed binary outcomes.
      - "Hormonal Contraceptives (years)" illustrated the duration of contraceptive usage among individuals.

- The box plot showed some variations in contraceptive usage duration based on "Schiller" and "Hinselmann", but distinctions weren't as clear as with the "Age" feature.

Comparison:

- The model with features "Schiller", "Hinselmann", and "Age" outperformed the model with "Schiller", "Hinselmann", and "Hormonal Contraceptives (years)" in terms of accuracy.
- Visual evidence from scatter plots and distribution graphs indicated that the combination with "Age" provided clearer distinctions in predicting biopsy outcomes.

Recommendation:

Given the observed accuracy scores and insights from the visualizations, we recommend using the feature set "Schiller", "Hinselmann", and "Age" for building a predictive model for users. This combination provides a more accurate and clearer distinction for predicting the target variable "Biopsy".

Further considerations:

- For real-world applications, always consider the ease of data collection. Age data might be more straightforward and more readily available than data on the duration of contraceptive usage.
- It's also beneficial to factor in the interpretability of the features when presenting results or making decisions based on the model's predictions.

Predictive Analysis Report on Cervical Cancer Dataset

Conclusion for the Predictive Analysis on Cervical Cancer Dataset

Through our comprehensive exploration and analysis of the cervical cancer dataset, our primary objective was to determine the optimal combination of features that would predict the target variable "Biopsy" most effectively.

1. Feature Importance:

- We identified the significance of various features using Random Forest's inherent capability to rank feature importances.

- The features "Schiller", "Hinselmann", and "Age" emerged as the top predictors, followed closely by "Hormonal Contraceptives (years)".

2. Model Evaluation:

- A predictive model using "Schiller", "Hinselmann", and "Age" achieved an impressive accuracy of approximately 96.41%.

- On the other hand, a model using "Schiller", "Hinselmann", and "Hormonal Contraceptives (years)" demonstrated an accuracy of 94.02%.

3. Visual Analysis:

- Visualizations highlighted the clear distinctions in biopsy outcomes, especially when considering the "Age" feature in combination with "Schiller" and "Hinselmann".

- The continuous features, "Age" and "Hormonal Contraceptives (years)", provided insights into their distributions and their potential relationships with the binary outcomes of "Schiller" and "Hinselmann".