**Project Analysis Report: Skin Disorder Prediction**

**1. Introduction:**

The objective of this project is to create a predictive model using machine learning techniques to identify various classes of skin diseases and provide suggestions to doctors for early diagnosis. The dataset contains 34 independent features, and the target variable represents different skin diseases, with disease 1 being the most frequently occurring and disease 6 being the least frequent.

**2. Data Preprocessing:**

- During the univariate analysis, it was observed that none of the features were normally distributed, and several features exhibited positive skewness. To handle missing values in the "Age" column, '?' marks were replaced with NaN, and then the mean age was imputed for these missing values.

- Nine independent features were found to have high correlation with multiple other features, leading to multicollinearity. These features were dropped to mitigate the issue of multicollinearity.

**3. Feature Engineering:**

- The heatmap analysis revealed several pairs of highly correlated features. For instance, "polygonal\_papules" showed strong correlations with "melanin\_incontinence," "vacuolisation\_and\_damage\_of\_basal\_layer," "saw-tooth\_appearance\_of\_retes," and "band-like\_infiltrate."

**4. Model Building and Evaluation:**

- Six different models were trained on the preprocessed data: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes.

- The initial Decision Tree and Random Forest models exhibited signs of overfitting, which were gone through hyperparameter tuning using GridsearchCV and RandomizedSearchCV.

- SVM model was found to be underfitting the data, while KNN showed average performance on both test and training data.

- Random Forest with Hyperparameter tuning achieved the best accuracy score of 95% on the test data and demonstrated better generalization compared to other models.

**5. Challenges Faced:**

- The dataset's high dimensionality (34 features) posed the challenge of the curse of dimensionality, potentially affecting model performance.

- Interpretation of the heatmap for feature correlation was difficult due to the large number of features.

- Pairplot visualization for multivariate analysis took significant time, and interpreting the small plots was challenging. Due to large number of plots in every single row.

- GridSearchCV for tuning the Random Forest model was time-consuming; hence, RandomizedSearchCV was employed for faster parameter optimization.

**6. Conclusion:**

After conducting a thorough analysis of various machine learning models, it was determined that the Random Forest model with hyperparameter tuning provides the best performance for production use. Both Naive Bayes and Random Forest with Hyperparameter tuning showed similar accuracy and F1-score on the test data, but the Random Forest model's higher training score and better generalization made it the preferred choice.

**7. Recommendations:**

- To further improve the model's performance, consider exploring more about Bagging and Boosting

- Investigate other algorithms or ensemble techniques that may better handle high-dimensional data and provide better generalization.

**8. References:**

1. [https://www.frontiersin.org/articles/10.3389/fmed.2020.00091/full]

2. [https://ieeexplore.ieee.org/document/9278810]