Project 2: Audience Questions and Answers – Milestone 3

Shanthibooshan Subramanian

Bellevue University

DSC680

Amirfarrokh Iranitablob

Here are ten questions that an audience might ask regarding the analysis of Water Potability.

1. How did you handle the imbalance in the dataset between potable and non-potable water samples, and why did you choose ADASYN for oversampling?

   ADASYN was chosen to handle class imbalance because it generates synthetic samples for the minority class, thus improving model performance and balance. This approach was selected over other methods like SMOTE due to its focus on generating samples near difficult-to-classify instances, which helps improve the model's sensitivity to the minority class.

2. Can you elaborate on the specific features or parameters that significantly impacted predicting water potability?

   Features such as turbidity, pH, and hardness were crucial in predicting water potability. These parameters significantly affected model predictions because they directly relate to water quality indicators that influence potability.

3. What challenges did you encounter during data preprocessing, and how did these impact the final model's performance?

   Challenges included handling missing values and normalizing features. These issues impacted the model's performance by potentially introducing biases or inconsistencies. Techniques like imputation and feature scaling were applied to mitigate these effects.

4. Were any specific geographic or seasonal trends in water potability that stood out from your analysis?

   Analysis revealed that water potability varied by region and season, with specific trends indicating higher contamination in certain geographic areas or during times of the year. This information can help in targeted water quality management and policy-making.

5. Considering your ethical implications, what measures did you take to ensure fairness and transparency in your model's predictions?

   Measures included ensuring data privacy, mitigating bias in the model, and providing transparency in model predictions. Regular audits and fairness checks were conducted to ensure the model's decisions were equitable and unbiased.

6. What are the most critical next steps for further improving the accuracy and applicability of predictive models in water quality management?

   Critical next steps include incorporating additional features (e.g., real-time environmental data), refining hyperparameters, and implementing ensemble methods. Continuous monitoring and periodic updates to the model are also essential to maintain accuracy.

7. What criteria did you use to select the machine learning algorithms for your analysis, and why were these chosen over others?

   Algorithms were selected based on their performance in handling imbalanced data, accuracy, and interpretability. Random Forest and XGBoost were favored for their robustness and ability to handle complex relationships in the data.

8. Can you elaborate on why Random Forest outperformed other models like SVM and KNN regarding accuracy and precision for predicting water potability?

   Random Forest outperformed due to its ensemble approach, which reduces overfitting and improves generalization. Its ability to aggregate results from multiple decision trees enhances accuracy and precision compared to models like SVM and KNN.

9. Given the dynamic nature of water quality and environmental factors, how do you plan to validate and update your predictive models over time?

   To address dynamic changes, the plan includes periodic retraining with new data, continuous model evaluation, and incorporating real-time data updates. This approach ensures the model remains accurate and relevant over time.

10. What were the key challenges you encountered during model deployment and how did you address them, particularly with the Random Forest model?

    Key challenges included scaling the model for production and integrating it with existing systems. These were addressed by optimizing model performance for deployment and ensuring robust integration with data pipelines and monitoring tools.