Discussion of results:  
A. Data Preprocessing

In our data preprocessing phase, several key steps were taken to address various issues and enhance the quality of our dataset. Firstly, we encountered missing values across the dataset, which we managed by replacing them with 'nan' for consistency and further handling. Additionally, within the 'MOSQUITO\_NUMBER' column, we encountered values reported in a range format. To ensure uniformity, we divided our approach into two subsets: one where we took the median of these ranges and another where we dropped these rows entirely. We found that the dataframe utilizing the median of the ranges yielded a higher accuracy of 71%, compared to 68% without this manipulation. Consequently, we adopted the former dataframe for our analysis. Furthermore, inconsistencies were observed in the 'INSECTICIDE\_CONC' column, where values were reported in 'µg.' To maintain consistency, these rows were dropped from consideration. Lastly, recognizing the imbalanced nature of our data, we implemented four methods: oversampling, a combination of oversampling and undersampling, stratified sampling, and stratified sampling coupled with oversampling and undersampling. These strategies aimed to mitigate the class imbalance issue, thereby enhancing the robustness of our models and ensuring more reliable predictions.

Using GridSearch to find the best decision tree, we observed the following values for the four different methods.

| **Method** | **Accuracy** | **F1 score** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| **Oversampling** | 0.636188 | 0.635582 | 0.638187 | 0.636487 |
| **Oversampling and Undersampling** | 0.721118 | 0.554239 | 0.594509 | 0.562154 |
| **Stratified Sampling** | 0.701495 | 0.548793 | 0.578368 | 0.552711 |
| **Stratified Sampling Coupled with Oversampling and Undersampling** | 0.697839 | 0.697597 | 0.697919 | 0.697846 |

| **Method** | **Accuracy** | **F1 score** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| **Oversampling** | 0.636188 | 0.635582 | 0.638187 | 0.636487 |
| **Oversampling and Undersampling** | 0.721118 | 0.554239 | 0.594509 | 0.562154 |
| **Stratified Sampling** | 0.701495 | 0.548793 | 0.578368 | 0.552711 |
| **Stratified Sampling Coupled with Oversampling and Undersampling** | 0.697839 | 0.697597 | 0.697919 | 0.697846 |