**Classification Algorithms – Scikit-learn Scavenger Hunt**

In Scikit-learn documentation you'll find a large list of ways to classify data. Do a little scavenger hunt in these docs: your goals is to look for classification methods and match a dataset in this curriculum, a question you can ask of it, and a technique of classification. Create a spreadsheet or table in a .doc file and explain how the dataset would work with the classification algorithm.

| **Algorithm** | **Dataset** | **Question You Could Ask** | **Explanation** |
| --- | --- | --- | --- |
| **Logistic Regression** | Pumpkin Dataset | Can we classify the pumpkin variety (e.g., Pie vs Decorative) based on features like price, city, and package? | Logistic Regression is a good baseline classifier. After encoding categorical features (city, package) and standardizing numerical ones (price, weight), the model can predict pumpkin variety as a binary or multi-class classification task. |
| **K-Nearest Neighbors (KNN)** | Iris Dataset (load\_iris) | What species of iris flower is this, based on sepal/petal measurements? | KNN works by comparing new data points to the closest training samples. Since Iris features are continuous (length/width), KNN can accurately classify new samples into setosa, versicolor, or virginica. |
| **Support Vector Machine (SVM)** | Digits Dataset (load\_digits) | What digit (0–9) is shown in this grayscale image? | SVM finds the best hyperplane to separate classes in high-dimensional space. For the digits dataset, each image (8x8 pixels) is transformed into features. SVM can then classify handwritten digits with strong performance. |
| **Decision Tree Classifier** | Titanic Dataset (or a similar CSV with passenger info) | Will a passenger survive the Titanic disaster based on age, class, and sex? | Decision Trees split data based on feature thresholds (e.g., if sex=female, survival odds increase). They are interpretable and handle both categorical and numeric variables. This makes them great for survival classification tasks. |
| **Naive Bayes (GaussianNB)** | Breast Cancer Dataset (load\_breast\_cancer) | Is this tumor malignant or benign based on medical imaging features? | Naive Bayes assumes independence between features. In the breast cancer dataset, features such as cell radius, texture, and smoothness are used. Despite its simplicity, Naive Bayes is often very effective for medical classification. |

This table shows how different classification algorithms in Scikit-learn align with datasets you’ve studied or that are widely used for learning. Each dataset naturally fits with specific algorithms based on the type of data (numerical, categorical, image-like). Together, they demonstrate the versatility of classification methods in practical scenarios.