|  |  |
| --- | --- |
| NAME: RAHUL THAMBI | ROLL NUMBER: A059 |
| SUBJECT: DWM | PROJECT-2 |

**PROJECT PART-2 (weblinks)**

**MATERIALS REFERRED TO FOR FINDING THE ALGORITHMS FOR THE SPECIFIED DATA MINING TASK:**

* The scikit-learn library provides a wide range of machine learning algorithms for data mining tasks, including classification, regression, clustering, and dimensionality reduction. <https://scikit-learn.org/stable/index.html>.
* The Weka software is a popular tool for data mining and machine learning. It provides a graphical user interface for exploring datasets and applying machine learning algorithms, including decision trees, association rule mining, and clustering. <https://www.cs.waikato.ac.nz/ml/weka/>.
* The KDnuggets website provides a wealth of resources for data mining, including tutorials, articles, and a directory of software tools and algorithms.

<https://www.kdnuggets.com/>.

**Algorithms identified:**

Given this dataset, there are several classification algorithms that you can use to create an accurate model that will determine if a person survived or not. Some of the commonly used algorithms are:

Logistic Regression: This is a simple and effective algorithm that can be used for binary classification problems, where the target variable has two classes (in this case, survived or not survived).

Random Forest: This is an ensemble learning algorithm that combines multiple decision trees to produce a more accurate model. Random Forest can handle non-linear relationships between features and the target variable.

Support Vector Machines (SVM): This algorithm creates a hyperplane that separates the data into two classes. SVM can handle non-linear decision boundaries by using a kernel function.

K-Nearest Neighbors (KNN): This algorithm classifies a data point by finding its k nearest neighbors in the training set and taking the majority vote of their classes.

Naive Bayes: This is a probabilistic algorithm that assumes that the features are independent of each other. Naive Bayes can be trained quickly and is often used for text classification problems.

Gradient Boosting: This is an ensemble learning algorithm that combines multiple weak classifiers to produce a strong classifier. Gradient Boosting can handle non-linear relationships between features and the target variable.

Neural Networks: This is a powerful algorithm that can learn complex non-linear relationships between features and the target variable. Neural Networks require a large amount of data and can be computationally expensive.

**Algorithms Shortlisted and why?**

**Naive Bayes algorithm:** is a probabilistic algorithm that is used for classification problems. It is based on the Bayes' theorem, which states that the probability of a hypothesis (in this case, the class of an instance) is proportional to the probability of the evidence (the values of the features) given that hypothesis. In the context of the Titanic dataset, Naive Bayes algorithm can be used to predict whether a passenger survived or not based on the available features suc as age, sex, ticket class, etc. Naive Bayes assumes that the features are independent of each other, which means that the presence or absence of a particular feature does not affect the probability of the presence or absence of any other feature. This assumption may not hold true in all cases, but it simplifies the computations and makes the algorithm faster and more efficient.

**REP Tree (Repeated Incremental Pruning to Produce Error Reduction Tree):** is a decision tree algorithm that is commonly used in machine learning for classification and regression problems. The algorithm works by recursively partitioning the dataset into smaller subsets based on the values of the input features and creating a tree structure that represents the decision process. Each node of the tree corresponds to a decision based on the value of a feature, and each leaf node corresponds to a predicted class or value. With the Titanic dataset, REP Tree could contribute to your analysis by providing a decision tree model that predicts whether a passenger survived or not based on their characteristics. The model could help identify the most important features that contribute to survival, such as gender, age, ticket class, and cabin location. Additionally, REP Tree could help identify any interactions or nonlinear relationships between features that affect survival. The output of the model could be used to gain insights into the factors that influenced survival and to make predictions for new data based on the learned decision rules.

**The LWL (Locally Weighted Learning) algorithm:** is a type of instance-based learning algorithm that can be used for classification problems. The LWL algorithm works by assigning weights to each training example based on how close it is to the test example being classified. The algorithm then uses these weights to calculate a weighted average of the class labels of the training examples to predict the class label of the test example. In the case of the Titanic dataset, LWL algorithm could be used as a classification algorithm to predict whether a passenger survived or not based on their demographic information, ticket information, and cabin information.

**The PART algorithm** is a decision tree algorithm that generates a set of rules from a given dataset to classify new data points. It is similar to the C4.5 algorithm but is designed to handle continuous and nominal data types.The PART algorithm can be used to classify the Titanic dataset by creating a decision tree that uses the attributes of the passengers to determine whether they survived or not. The algorithm will partition the data into subsets based on the values of the attributes, and recursively create rules that predict the class of a new data point based on its attribute values.

**Logistic regression:** is a widely used classification algorithm that can be applied to the Titanic dataset to predict whether a passenger survived or not based on several features. Logistic regression is particularly useful for binary classification problems like this, where the target variable has only two possible outcomes (survived or not survived). The logistic regression algorithm models the relationship between the input features and the probability of the target variable. It estimates the probability of a passenger surviving based on the values of the input features, such as age, gender, ticket class, and others.