**GPU Vs. CPU data report- Hagay Ringel (June 02, 2023)**

Introduction:

In this project, a dataset from Kaggle containing technical information about GPUs and CPUs was analyzed. The dataset includes details such as the number of transistors, frequency, process size, as well as categorical information like product name, vendor, and type. The analysis aims to achieve two main goals:

1. Provide a comprehensive understanding of trends, patterns, and an overview of the data through visualization techniques.
2. Build a model to predict the manufacturer based on technical factors, specifically the number of transistors and die size. This correlation can significantly impact important business decisions related to supplier selection, quality control, cost optimization, and more.

Methodology:

The dataset was prepared by handling null values, employing two techniques: dropping unnecessary columns or those with excessive null values, and replacing null values with the average value of the respective column.

Data Visualization:

Visualizations were utilized to identify patterns and uncover insights not easily noticeable to the human eye. The distribution of items among foundries was displayed using a pie chart. A scatter plot was employed to showcase the relationship between frequency and transistor quantity, categorized by vendor and type (GPU/CPU). Another visualization focused on validating Moore's Law (Moore's Law still holds, especially in GPUs). Various other valuable visuals were generated to enhance the understanding of the data.

Machine Learning Model - Random Forest:

To predict the manufacturer based on the given features, the Random Forest algorithm was chosen for the following reasons:

1. Improved Performance: Random Forest generally outperforms individual decision trees, especially for complex or high-dimensional datasets.
2. Feature Importance: Random Forest provides measures of feature importance, aiding in understanding the relative contribution of different features to the classification task. This is crucial since two features were utilized in this analysis.
3. Versatility: Random Forest can be applied to various types of tasks, including classification, which aligns with the goal of this model.

The dataset was split into a training set (70%) and a testing set (30%). Predictions were made using the testing set, and a confusion matrix was generated to evaluate the model's performance. Finally, a classification report was created, summarizing the model's outcomes. According to our model these are the confusion matrix results:

A picture containing text, screenshot, software, display

Description automatically generated

Precision: Precision measures the proportion of correctly predicted instances among the instances predicted as a specific class. The highest precision is achieved for Other Vendor (1.00), indicating that all predictions for this class were correct. Intel and AMD also have high precision values (0.97 and 0.98, respectively), indicating a low rate of false positives. NVIDIA and ATI have slightly lower precision values of 0.93 and 0.91, respectively.

Recall: Recall measures the proportion of correctly predicted instances of a class among all instances of that class in the dataset. The classes with the highest recall are ATI (0.98), Intel (0.97), and NVIDIA (0.96). AMD has a slightly lower recall value of 0.93. Other Vendor has the lowest recall of 0.88, suggesting that some instances of this class were misclassified as other classes.

F1-score: F1-score is the mean of precision and recall, providing a single metric that balances both measures. The weighted average F1-score is high for all classes, ranging from 0.93 (Other Vendor) to 0.97 (Intel).

Accuracy: The overall accuracy of the model is 0.95, indicating that 95% of the instances in the dataset were correctly classified.

Overall, the model performs well with high precision, recall, and F1-score values for most classes. However, it has a lower recall for the Other Vendor class, indicating that it struggles to correctly identify instances of this class. It would be valuable to investigate why the model has lower performance for this specific class and potentially explore strategies to improve its classification accuracy for the Other Vendor category.

Note: This project serves as an example of a model used to classify vendors based on technical factors. However, it can be easily adapted to assess other technical factors for predicting the vendor or to predict a different variable, such as the type. The significance of this project lies in showcasing the utilization of visualization tools and machine learning techniques to derive business insights and make data-driven decisions.