**Documentation Report**

**Gender Classification using Machine Learning**

This task involved training various machine learning models by using all estimators on a dataset that comprised samples for both male and female genders.

**Data Collection and Preparation:**

The dataset used for this task was collected from Kaggle that contains images of male and female faces. It consisted of a balanced number of samples for each gender, with 23000 samples per gender category but due to large data we train models only on 1200 data of each category to reduce computation cost.

**Model Selection and Experimentation:**

To find the most suitable model for gender classification, I experimented with a range of machine learning algorithms using all estimators. The models considered included:

| **Model Name** | **Precision** |
| --- | --- |
| AdaBoostClassifier | 0.759494 |
| BaggingClassifier | 0.830769 |
| BernoulliNB | 0.632867 |
| CalibratedClassifierCV | 0.745614 |
| CategoricalNB | 0.632867 |
| ComplementNB | 0.674797 |
| DecisionTreeClassifier | 0.603239 |
| DummyClassifier | 0.483333 |
| ExtraTreeClassifier | 0.642276 |
| ExtraTreesClassifier | 0.800905 |
| GaussianNB | 0.625430 |
| GaussianProcessClassifier | 0.701550 |
| GradientBoostingClassifier | 0.806167 |
| HistGradientBoostingClassifier | 0.833333 |
| KNeighborsClassifier | 0.705455 |
| LabelPropagation | 0.000000 |
| LabelSpreading | 0.000000 |
| LinearDiscriminantAnalysis | 0.635593 |
| LinearSVC | 0.728814 |
| LogisticRegression | 0.742489 |
| LogisticRegressionCV | 0.810573 |
| MLPClassifier | 0.807175 |
| MultinomialNB | 0.674797 |
| NearestCentroid | 0.628472 |
| NuSVC | 0.866071 |
| PassiveAggressiveClassifier | 0.755760 |
| Perceptron | 0.737991 |
| QuadraticDiscriminantAnalysis | 0.509579 |
| RandomForestClassifier | 0.809091 |
| RidgeClassifier | 0.628571 |
| RidgeClassifierCV | 0.688034 |
| SGDClassifier | 0.816327 |
| SVC | 0.857778 |

**Model Evaluation and Results:**

After conducting thorough experiments and evaluations, the Support Vector Classifier (SVC) emerged as the standout model for this task. The performance metrics achieved by the SVC model were as follows:

* Accuracy: 0.852
* F1 Score: 0.851
* Precision: 0.877

These results indicate that the SVC model performed exceptionally well in accurately classifying gender based on the provided data. The high accuracy, F1 score, and precision scores demonstrate the effectiveness of this model for this specific classification task.

**Challenges and Limitations:**

* **Handling Images with Machine Learning:** One of the major challenges faced during this project was the handling of image data using traditional machine learning techniques. Images are inherently high-dimensional data, and processing them can be computationally expensive and memory-intensive. This can lead to challenges in feature extraction, model training, and inference.
* **Low Accuracy:** Although the SVC model achieved a relatively high accuracy of 0.852, it's worth noting that image-based gender classification is a complex task. Factors such as variations in lighting, facial expressions, and age can significantly impact the accuracy of the model. Achieving even higher accuracy can be challenging without more extensive data and more advanced techniques, such as deep learning.
* **Image Resizing and Computational Time:** To reduce computational complexity, image resizing was performed as a preprocessing step. However, this resizing process significantly extended the time required for data preparation. Balancing computational efficiency with model accuracy is an ongoing challenge in image-based tasks.
* **Data Reduction and Impact on Accuracy:** To improve computational efficiency, a reduction in the amount of data was performed. While this helped with computation, it also had an impact on the model's ability to generalize and achieve higher accuracy.

**Conclusion:**

In conclusion, the successful completion of the gender classification task using the SVC model highlights both the potential and limitations of machine learning for image-based classification tasks. While the model achieved respectable accuracy, it's important to acknowledge the complexities of working with image data, the challenges associated with achieving higher accuracy, and the trade-offs in computational efficiency.

Future work in this area could involve exploring deep learning approaches, optimizing image preprocessing steps, and seeking a balance between data reduction for computation and model accuracy.