Team Members:

1. Mohammed Ashraf Mohammed 20191700509

2. Mostafa Labib 20191700646

3. Mostafa Ayman Awad 20191700640

4. Mostafa Ashraf 20191700639

5. Aisha Soliman 20191700330

6. Mostafa Hesham 20191700656

News group classification

Project Title

Introduction:

The main idea of the project is to classify documents given to you into one of 20 categories by training a model on the given training set.

Methodology:

In order to classify documents in the test set, we had to prepare the train data. The data preparation tasks were as follows:

* Convert all letters to lowercase
* Word tokenization
* Stop words removal
* Lemmatization
* Feature Extraction using TfIdfVectorizer

After the data was ready to be input to a model, we tried 3 different models as follows:

* Logistic regression (linear)
* SVM (using poly kernel with degree 2)
* Decision tree

Data Set Summary:

1-What is the data set used?

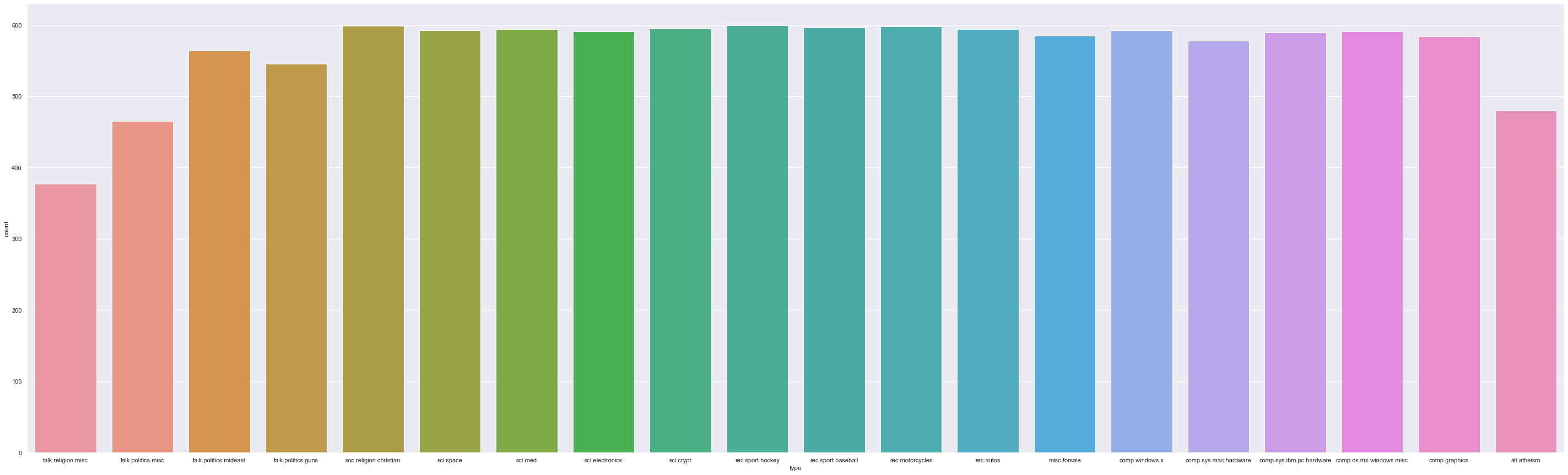
* [20news-bydate.tar.gz](http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz)

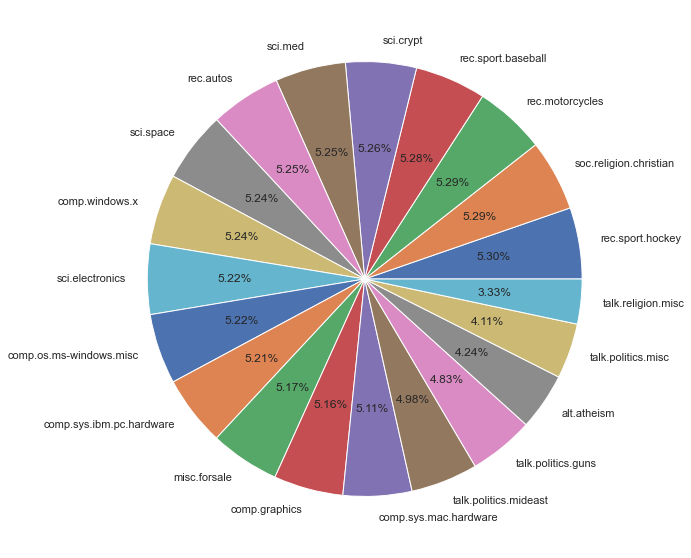
2- What is the summary of the dataset columns?

* Content: the document text.
* Name: id of the document in its category i.e. name of the document file.
* Type: news group type (target feature/label).

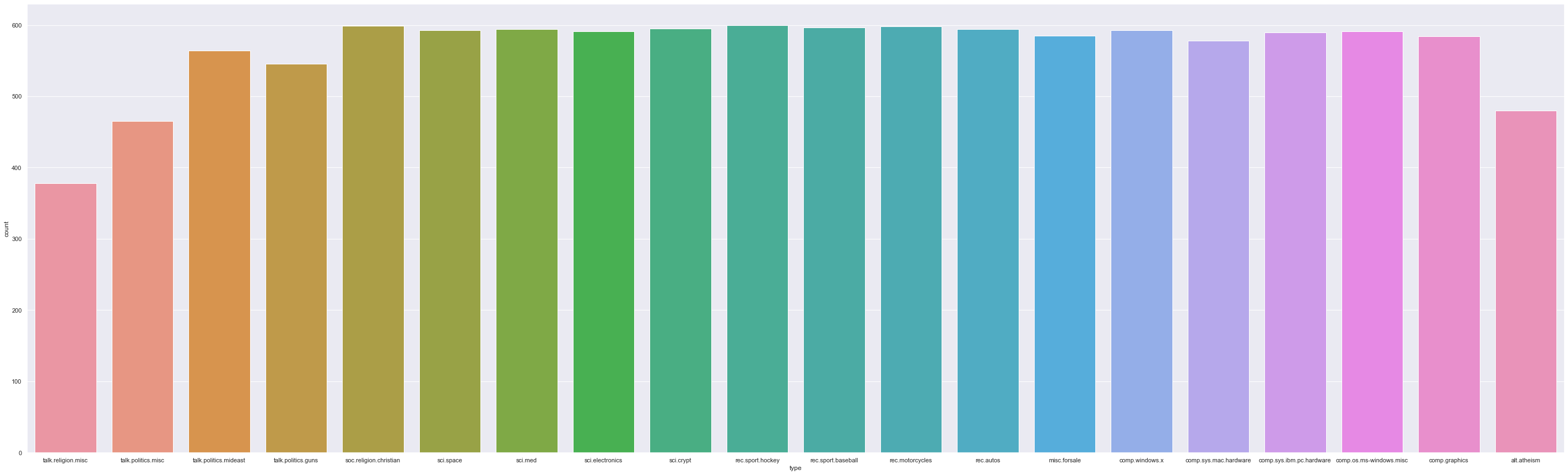
3- Visualize the dataset statistics

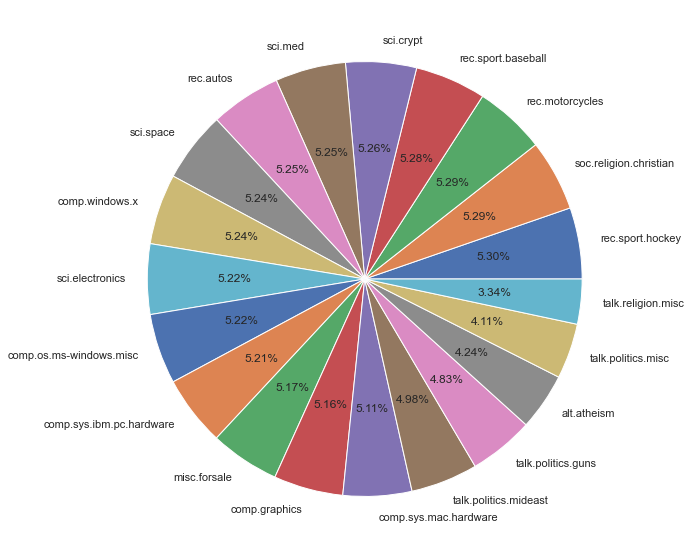
Training dataset:





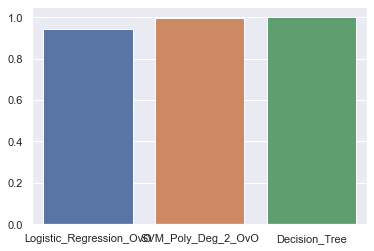
Test dataset:





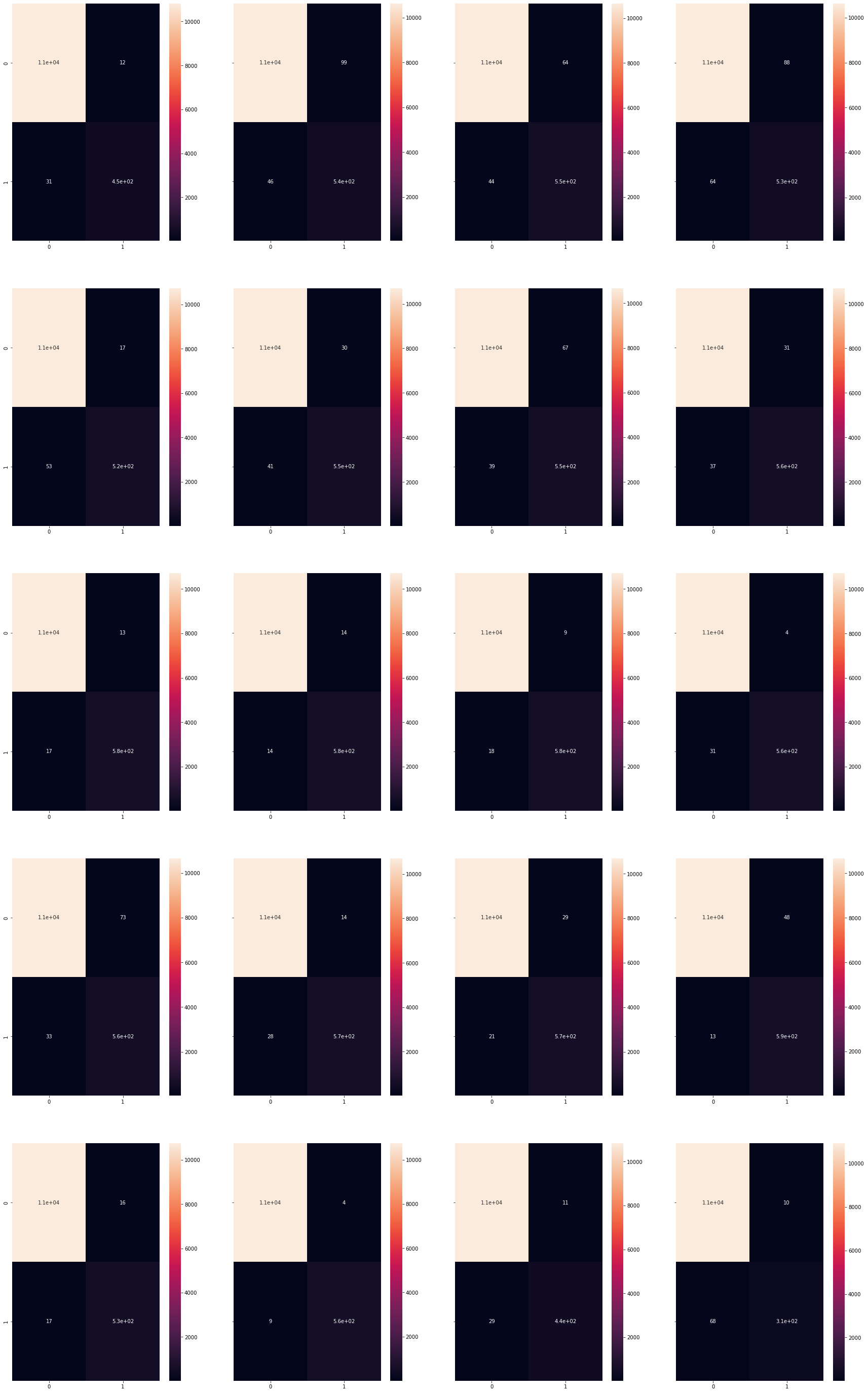
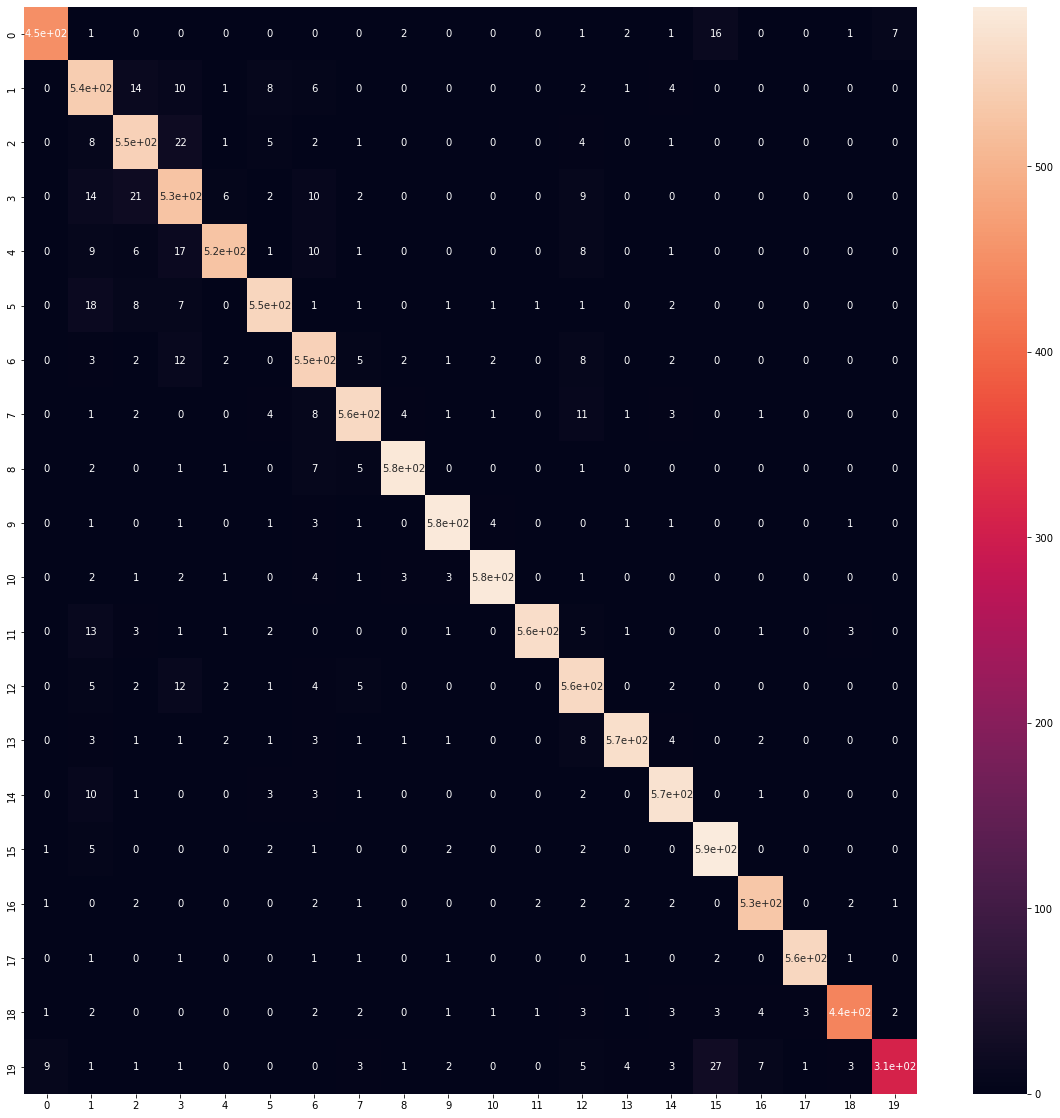
Results:

Training Accuracies

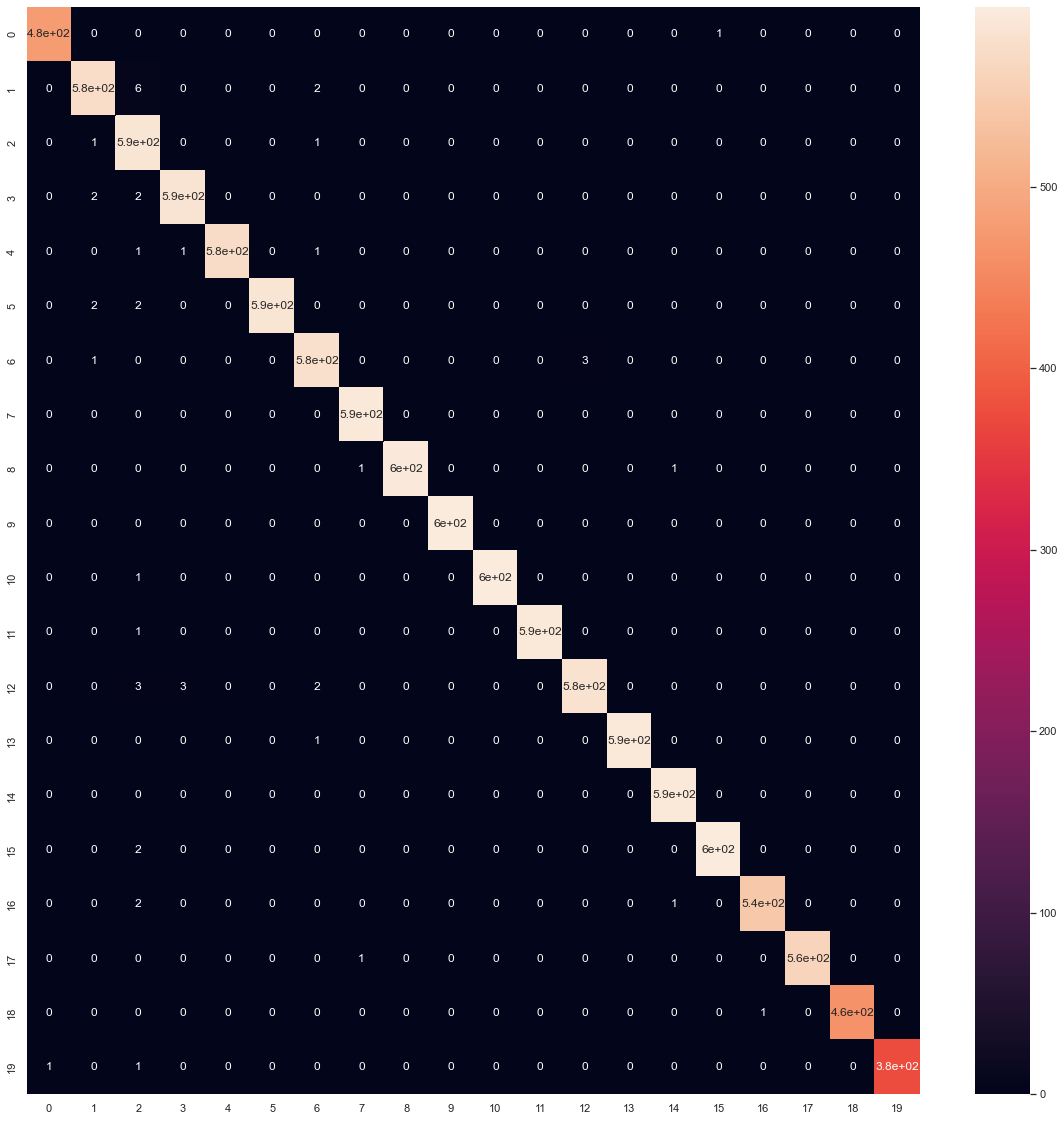


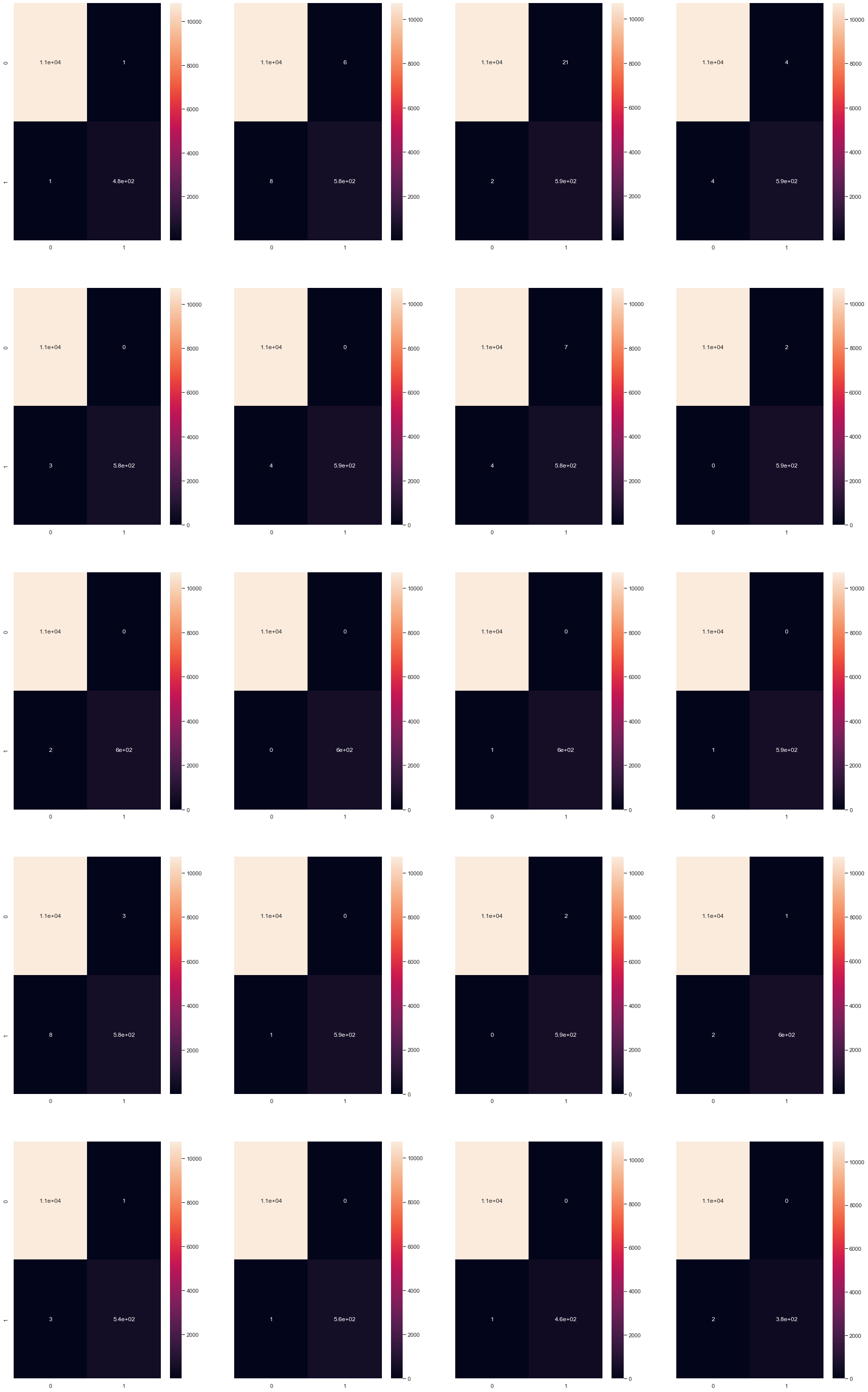
Testing Accuracies

Logistic Regression Model Confusion Matrix



SVM(Kernel: Polynomial,degee:2) Regression Model Confusion Matrix





Decision Tree Regression Model Confusion Matrix

