Hypothesis/Research Question:

This study aims to evaluate the performance of logistic regression on datasets encrypted using the BFV and CKKS homomorphic encryption schemes. Initially, logistic regression will be applied to the unencrypted (plaintext) datasets to establish baseline performance metrics. Subsequently, the same analysis will be conducted on the encrypted datasets using BFV and CKKS. The results will be compared based on key evaluation metrics, including F1-score, accuracy, precision, and recall. However, this time around it will be on genomic datasets and their indicators.

This data collection is to answer the following research question:

- **Can logistic regression be effectively performed on genomic datasets encrypted with BFV and CKKS?**
- **How closely do the performance metrics of encrypted logistic regression align with those of the plaintext baseline?**

Process:

The initial phase of this project involves creating two separate Python scripts, each designed to perform logistic regression on a plaintext dataset. One script will focus on implementing the BFV encryption scheme, while the other will implement the CKKS scheme. Following this, I will manually develop a logistic regression model based on the TenSEAL tutorial.

Once the encrypted logistic regression models are implemented using both BFV and CKKS, I will run each model on the corresponding encrypted dataset. Furthermore, I will change the BFV and CKKS parameters to measure if there are any discrepancies when security is increased. Finally, I will evaluate and compare the results by collecting performance metrics including accuracy, precision, recall, and F1-score, and generating classification reports for both the baseline (plaintext) model and the encrypted models.

The three datasets being utilized are datasets I created from the XENA database. I read papers to see which genomic indicators are the best to predict solid normal tissue or primary tumors for prostate cancer, breast cancer, and lung cancer. The list of indicators are the following:

Breast Cancer:

- BRCA1 - ENSG00000012048.19
- BRCA2 - ENSG00000139618.14
- LAMA2 - ENSG00000196569.11
- TIMP4 - ENSG00000157150.4
- TMTC1 - ENSG00000133687.15
- ESR1 - ENSG00000091831.21

- ESR2 - ENSG00000140009.18
- HER2 - ENSG00000141736.13
- PR - ENSG00000082175.14
- GATA3 - ENSG00000107485.15

Lung Squamous Cell Carcinoma
- TP53 - ENSG00000141510.15
- GRM8 - ENSG00000179603.17
- BAI3 - ENSG00000163682.15
- ERBB4 - ENSG00000178568.13
- RUNX1T1 - ENSG00000159216.18
- KEAP1 - ENSG00000079999.13
- FBXW7 - ENSG00000109670.13
- KRAS - ENSG00000133703.11
- SOX2 - ENSG00000181449.3
- FGFR1 - ENSG00000077782.19

Prostate Cancer:
- SPOP - ENSG00000145041.15
- ERG - ENSG00000157554.18
- PTEN - ENSG00000171862.9
- TP53 - ENSG00000141510.15
- MYC - ENSG00000136997.14
- AR - ENSG00000169083.15
- RB1 - ENSG00000139687.13
- ETS1 - ENSG00000134954.14
- ETS2 - ENSG00000157557.11

Results:

Table 1: Prostate Cancer Dataset

		F1-Score	**Precision**	**Recall**	**Accuracy**
BFV (scaling factor = 100000)	**4096, [30,20,30]**	0.43	0.43	0.43	0.45
	8192, [30, 20, 30]	0.46	0.46	0.46	0.48
CKKS	**4096, [30,20,30], 2**20**	0.90	0.91	0.90	0.90
	8192, [30, 20, 30], 222**	0.87	0.87	0.87	0.89

Plain	**Not Applicable**	0.87	0.87	0.87	0.87

Table 2: Breast Cancer Dataset

		F1-Score	Precision	Recall	Accuracy
BFV (scaling factor = 100000)	**4096, [30,20,30]**	0.40	0.40	0.41	0.40
	8192, [30, 20, 30]	0.58	0.58	0.58	0.58
CKKS	**4096, [30,20,30], 2**20**	0.93	0.93	0.93	0.93
	8192, [30, 20, 30], 222**	0.90	0.90	0.90	0.90
Plain	**Not Applicable**	0.90	0.90	0.90	0.90

Table 3: Lung Cancer Dataset

		F1-Score	Precision	Recall	Accuracy
BFV (scaling factor = 100000)	**4096, [30,20,30]**	0.40	0.40	0.40	0.40
	8192, [30, 20, 30]	0.37	0.37	0.37	0.37
CKKS	**4096, [30,20,30], 2**20**	0.90	0.92	0.90	0.90
	8192, [30, 20, 30], 222**	0.91	0.91	0.91	0.91
Plain	**Not Applicable**	0.93	0.94	0.93	0.93

Other Notes:

The data shows similar results compared to the clinical datasets. The BFV encryption scheme is terrible, but the CKKS encryption scheme maintains the integrity of the metrics.