CSE6250: Big Data Analytics for Healthcare
Homework 1

Arindam Duttagupta
GT ID: 903327355

Q1) CITI Certification:

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Arindam Duttagupta (ID: 6883170)
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)
- **Institution Email:** aduttagupta3@gatech.edu

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 1 Biomedical research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 25823027
- **Completion Date:** 17-Jan-2018
- **Expiration Date:** 16-Jan-2021
- **Minimum Passing:** 70
- **Reported Score*:** 99

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
History and Ethics of Human Subjects Research (ID: 498)	16-Jan-2018	7/7 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	16-Jan-2018	5/5 (100%)
Informed Consent (ID: 3)	16-Jan-2018	5/5 (100%)
Social and Behavioral Research (SBR) for Biomedical Researchers (ID: 4)	16-Jan-2018	4/4 (100%)
Records-Based Research (ID: 5)	16-Jan-2018	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	16-Jan-2018	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	16-Jan-2018	5/5 (100%)
Research Involving Children (ID: 9)	16-Jan-2018	3/3 (100%)
Research Involving Pregnant Women, Fetuses, and Neonates (ID: 10)	16-Jan-2018	3/3 (100%)
International Studies (ID: 971)	16-Jan-2018	3/3 (100%)
FDA-Regulated Research (ID: 12)	16-Jan-2018	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	16-Jan-2018	5/5 (100%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	17-Jan-2018	4/4 (100%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	17-Jan-2018	4/5 (80%)
Avoiding Group Harms - U.S. Research Perspectives (ID: 14080)	17-Jan-2018	3/3 (100%)
Stem Cell Research Oversight (Part I) (ID: 13882)	17-Jan-2018	5/5 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k4aad8c70-7917-417d-abc7-5c6cb9898f5d-25823027

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

citiprogram.org/verify/?k4aad8c70-7917-417d-abc7-5c6cb9898f5d-25823027

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Arindam Duttagupta (ID: 6883170)
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)
- **Institution Email:** aduttagupta3@gatech.edu

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 2 Social / Behavioral Research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 25826001
- **Completion Date:** 17-Jan-2018
- **Expiration Date:** 16-Jan-2021
- **Minimum Passing:** 70
- **Reported Score*:** 93

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Students in Research (ID: 1321)	17-Jan-2018	4/5 (80%)
History and Ethical Principles - SBE (ID: 490)	17-Jan-2018	4/5 (80%)
Defining Research with Human Subjects - SBE (ID: 491)	17-Jan-2018	4/5 (80%)
The Federal Regulations - SBE (ID: 502)	17-Jan-2018	5/5 (100%)
Assessing Risk - SBE (ID: 503)	17-Jan-2018	5/5 (100%)
Informed Consent - SBE (ID: 504)	17-Jan-2018	5/5 (100%)
Privacy and Confidentiality - SBE (ID: 505)	17-Jan-2018	5/5 (100%)
Research with Children - SBE (ID: 507)	17-Jan-2018	4/5 (80%)
Research in Public Elementary and Secondary Schools - SBE (ID: 508)	17-Jan-2018	5/5 (100%)
International Research - SBE (ID: 509)	17-Jan-2018	5/5 (100%)
International Studies (ID: 971)	16-Jan-2018	3/3 (100%)
Internet-Based Research - SBE (ID: 510)	17-Jan-2018	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	16-Jan-2018	5/5 (100%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	17-Jan-2018	4/4 (100%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	17-Jan-2018	4/5 (80%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k9dccb615-5fa2-4daf-bfe1-34994d9c94e7-25826001

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

citiprogram.org/verify/?k9dccb615-5fa2-4daf-bfe1-34994d9c94e7-25826001

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Arindam Duttagupta (ID: 6883170)
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)
- **Institution Email:** aduttagupta3@gatech.edu

- **Curriculum Group:** CITI Health Information Privacy and Security (HIPS)
- **Course Learner Group:** CITI Health Information Privacy and Security (HIPS) for Biomedical Research Investigators
- **Stage:** Stage 1 - HIPS

- **Record ID:** 25823028
- **Completion Date:** 23-Jan-2018
- **Expiration Date:** 22-Jan-2021
- **Minimum Passing:** 70
- **Reported Score*:** 96

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Basics of Health Privacy (ID: 1417)	23-Jan-2018	4/5 (80%)
Health Privacy Issues for Researchers (ID: 1419)	23-Jan-2018	5/5 (100%)
Basics of Information Security, Part 1 (ID: 1423)	23-Jan-2018	5/5 (100%)
Basics of Information Security, Part 2 (ID: 1424)	23-Jan-2018	5/5 (100%)
Protecting Your Computer (ID: 1425)	23-Jan-2018	5/5 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?kf0d730e4-58a4-463f-a8f5-a747171196eb-25823028

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

citiprogram.org/verify/?kf0d730e4-58a4-463f-a8f5-a747171196eb-25823028

Q2) Descriptive Statistics:

Metric	Deceased Patients	Alive Patients	Function to complete
Event Count 1. Average Event Count 2. Max Event Count 3. Min Event Count	 982.0 15572 1	 498.0 8272 1	Event_count_metrics
Encounter Count 1. Average Encounter Count 2. Max Encounter Count 3. Min Encounter Count	 23 203 1	 15 391 1	Encounter_count_metrics
Record Length 1. Average Record Length 2. Max Record Length 3. Min Record Length	 43 1739 0	 44 2906 0	record_length_metrics

Q4) Predictive Modelling:

Table 2: Model Performance on the Training data

Model	Accuracy	AUC	Precision	Recall	F1-Score
Logistic Regression	0.9545	0.9454	0.9869	0.8988	0.9408
SVM	0.9940	0.9945	0.9882	0.9970	0.9925
Decision Tree	0.7763	0.7475	0.7921	0.6011	0.6835

Table 3: Model Performance on the given Validation Data

Model	Accuracy	AUC	Precision	Recall	F1-Score
Logistic Regression	0.7380	0.7375	0.6804	0.7333	0.7058
SVM	0.7380	0.7388	0.6767	0.7444	0.7089
Decision Tree	0.6714	0.6569	0.6329	0.5555	0.5917

Table 4: Cross Validation (Logistic Regression)

Cross Validation	Avg Accuracy	Avg AUC
K-Fold	0.7285	0.7115
Randomized	0.7357	0.7142

Q4.1d)

Based on the tests performed on features_svmlight. validate dataset, we see that there is a sharp decrease in performance metrics such as Accuracy and AUC, and the Decision Tree classifier gives the best results compared to Logistic Regression and SVM. The main reason for this performance drop could be overfitting, as it fails to generalize to previously unseen test data. Alternatively, the train and test data might not follow the same distribution, which results in drop in performance even without overfitting.

After experimenting with other classifiers, I found that the soft Voting classifier algorithm achieved the best AUC. The idea behind the Voting Classifier is to combine conceptually different machine learning classifiers and use the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses. Here we used Bagging classifier (to tackle overfitting) and Adaboost classifier (to reduce bias) as components of Voting classifier. Using 80 estimators for Bagging and 45 estimators for Boosting, I achieved an AUC score of 78.611% (a 4.73% improvement over the previous best result of SVM)

Q4.3b)

In my best predictive model, the features from etl.py are used for training the Voting classifier (Bagging and Adaboost). I also experimented with other classifiers present in the scikit-learn package, and found that this weighted combination of Bagging and Adaboost provides a top AUC score of 71.663% in Kaggle.