BLG553E Special Topics in Computer Science - Trustable AI

Course Description:

Machine Learning models and solutions based on them are increasingly operationalized. This course will cover how to evaluate, communicate, update trustworthiness of a machine learning solution. Trustworthiness will be evaluated under validity, privacy, explainability and responsibility components. AI interpretation methods will be covered in depth. Applications in different industries will be examined. Students will need prepare a project where they help an AI to become more trustworthy through one or more of the aspects above. Python programming and Machine Learning or an equivalent course are prerequisites.

Weeks	Topics (Tentative)
1	Introduction (What is Trustable AI and why is it important, regulations)
2	Valid AI
3	Privacy Preserving AI
4	Responsible AI
5	Interpretable AI Methods (1)
6	Interpretable AI Methods (2)
7	Interpretable AI – to Machine and Data Scientists
8	Interpretable AI – to Domain Expert
9	UI/UX for Trustable AI
10	Trustable AI in Finance & Insurance
11	Trustable AI in Healthcare
12	Trustable AI in Manufacturing and Infrastructure
13	Trustable AI in Education and Law
14	Project Presentations

Ödevler **(Homework)**	**4**	**50%**
Projeler **(Projects)**	**1**	**50%**