**Abstract**

Diabetes mellitus, also known as type 2 diabetes, is a significant problem in the UK and many other countries across the world. More than 4.3 million people in the UK currently live with the disease, with an estimated 850,000 who could be living with the disease but are yet to be diagnosed. This puts a significant strain on NHS services, with overall diabetes treatment totalling 10% of the NHS’ total budget.

Machine Learning can be utilised as a method for early screening and prevention of diabetes, which is demonstrated in this study using surveillance data from BRFSS 2021 survey. Diabetes risk was predicted using Logistic Regression, Decision Tree, Random Forest and Gradient Boosting models. Several model evaluation methods such as Area under the Receiver Operating Characteristic Curve (AROC), Accuracy, F1 Score and Fbeta Score. The XG Boost model was found to provide the highest accuracy (0.807), and joint-highest AROC with Logistic Regression (0.807). The Logistic Regression model was found to have the highest sensitivity (0.741) and Fbeta Score (0.595), which suggests that model is the most suitable model for the diabetes use case.

Opportunities for further work are discussed, including productionising a machine learning model in the form of the cloud solution and producing a dashboard to displaying the results of diabetes predictions. There is also discussion over the considerations that would need to be made from the use of PII in the event patient-level NHS data is used in an extension of this study.