Choose Your Own Project: Pancreatic ductal adenocarcinoma biomarkers

Daniel Flores Orozco

2021-12-16

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic neoplasm and is considered a highly aggressive and lethal cancer (Sarantis et al., 2020). In fact, less than 20% of individuals diagnosed with PDAC survive the first year, and less than 10% survived more than five years (Debernardi et al., 2020; Sarantis et al., 2020). The poor prognosis is mainly attributed to late diagnosis and the lack of efficient treatments during an early stage. However, when the disease is detected in the early stages, the probability of surviving more than five years increases up to 70% (Debernardi et al., 2020). Therefore, the development of novel tests for the early detection of PDAC is essential to improve the patients' prognosis.

Debernardi et al. (2020) identified a group of proteins (LYVE1, REG1A, REG1B, TFF1, and plasma CA19-9) in the urine that could potentially serve as biomarkers to detect PDAC. Debernardi et al. (2020) developed an algorithm to differentiate cancer samples from benign samples using LYVE1, REG1B, and TFF1 levels, as well as creatinine levels and age as predictors. Using logistic regression, they achieved a sensitivity and specificity of 0.963 and 0.967, respectively.

In this document, different Machine Learning algorithms, including random forest, kNN, SVM, logistic regression, and bayesian logistic regression, were used to predict if a given sample belonged to a benign or cancer group using the data provided by Debernardi et al. (2020). The ultimate objective of this project was to improve the performance reported by the original authors.

2. Methodology

2.1 Database

The pancreatic cancer biomarkers data was downloaded from the database repository Kaggle (https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer).

2.2 Data exploration

The first step was to explore the