Dataset Description: GSE208581

The dataset GSE208581 comprises gene expression profiles obtained from peripheral blood samples of patients who underwent major surgery. The study aimed to investigate the gene expression patterns associated with different postoperative outcomes, specifically focusing on the development of sepsis.

Sample Groups:
The samples are categorized into four distinct groups based on the patients' postoperative outcomes:

1. sep+: Patients who developed postoperative infection leading to sepsis.
2. SIR+: Patients who experienced a non-infectious systemic inflammatory response.
3. UInf+: Patients who developed an uncomplicated postoperative infection.
4. SIR-: Patients who had an uncomplicated postoperative course without significant inflammatory response.

Data Analysis:
For the analysis, samples specifically related to sepsis patients (Sepsis+) and those with an uncomplicated postoperative course (SIRS-) were extracted to identify gene expression changes associated with sepsis development.

_____

Presented-genes: 53
Missing-genes : S100A8 S100A9

_____

Data normalization:
Dataset was normalized because the result from the mRNA-count dataset was weird. Some columns had 0 values. Normalization has been done by vsn to prepare data for downstream analysis.

Random Forest Model:

I used the sepsis_vsn_dataGSE208581.csv dataset to apply Random forest which is for sepsis patients and Patients who had an uncomplicated postoperative course without significant inflammatory response(SIRS-) as control.
The dataset contains: [187 rows, 55 columns].

Here is the code that is used for both held-out validation and the leave-one-out validation (RF-labelGSE208581cose.R).

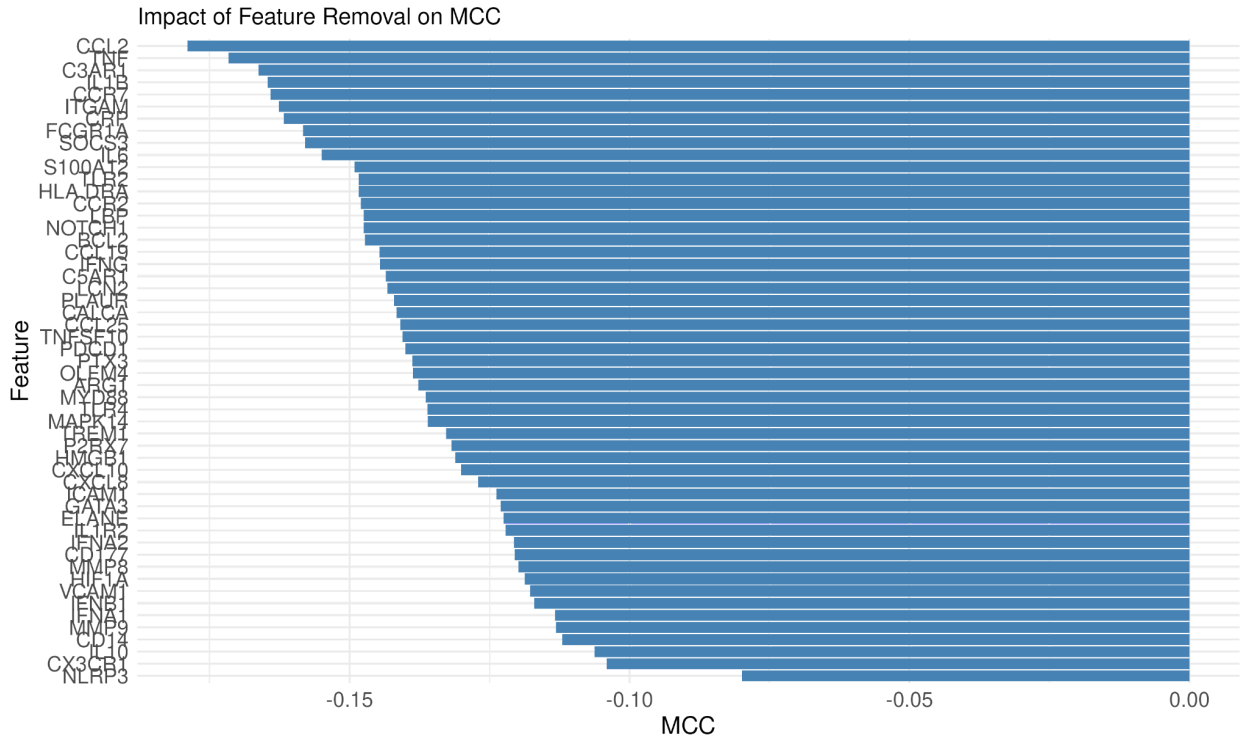
For the held-out validation that was repeated 100 times the result is here (100-times-repeated-split) and thea average of 100-time is: (average-metrics)

average_metrics

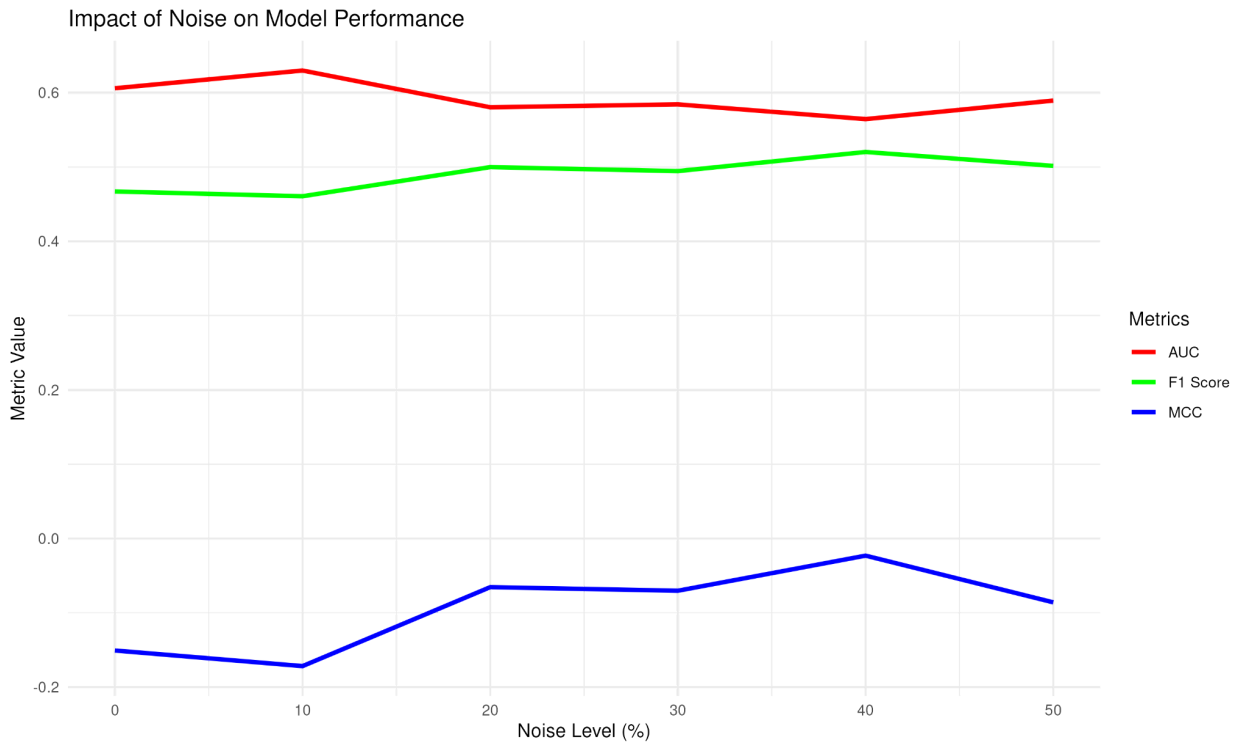
MCC	F1	AUC	TPR	TNR	PPV	NPV
-0.126117338682493	0.476337449062237	0.606075757575758	0.672666666666667	0.219090909090909	0.370294288890905	0.480226717726718

Feature Removal Analysis:

The feature removal results indicate the contribution of individual features (genes) to the model's performance. Metrics such as MCC, F1 Score, AUC, TPR, TNR, PPV, and NPV are used to evaluate the importance of each feature.



Sanity check:



The model is performing poorly, especially with MCC being negative or close to zero and F1 scores being low. This indicates issues with precision, recall, or the balance between classes (e.g., misclassification of the minority class).