Final Exam

Advanced Methods in Applied Statistics

Cyan Yong Ho Jo
bsf873

5th April 2024

I Cyan Yong Ho Jo expressly vow to uphold my scientific, academic, and moral integrity by working individually on this exam and soliciting no direct external help or assistance.

PROBLEM 1

a

Below are the functions that could potentially govern the distribution of the data. Instead of these functions, the data distribution may adhere to one of three discrete distributions: binomial, Poisson, or logarithmic. I compute the maximum likelihood values for each dataset across all functions and the three discrete distributions. The best-fitted function or distribution for each data is shown in TableI with the likelihood value. However, some likelihood values look strange, I can sense that there must have been a mistake in the calculation of likelihood for the functions.

$$
\begin{aligned}
f_1(a) &: x \mapsto \frac{1}{x+5}\sin(ax) \\
f_2(a) &: x \mapsto \sin(ax)+1 \\
f_3(a) &: x \mapsto \sin(ax^2) \\
f_4(a) &: x \mapsto \sin\left((ax+1)^2\right) \\
f_5(a) &: x \mapsto x\tan(ax) \\
f_6(a,b) &: x \mapsto 1+ax+bx^2 \\
f_7(a) &: x \mapsto 5+ax \\
f_8(a,b,c) &: x \mapsto \sin(ax)+c\exp(bx)+1 \\
f_9(a,b) &: x \mapsto \exp\left(-\frac{(x-a)^2}{2b^2}\right)
\end{aligned}
$$

b

The normalised distribution of each dataset with the normalised best fits. As the process of finding the best fits has issues, The fitting functions drawn here are not correct!

Column	**Best Fit**	**LLH**	**Parameters**
first	f_6	-14197.77	$a = 10.0$, $b = 10.0$
second	f_6	-3978.596	$a = 10.0$, $b = -8.384$
third	Poisson	9570.106	$\lambda = 6.0$
fourth	f_9	-5141.899	$a = 1.111$, $b = -0.101$
fifth	f_6	-3307.289	$a = 10.0$, $b = -2.727$
sixth	f_6	-4553.911	$a = 10.0$, $b = 10.0$

Table I: Best fit ln-Likelihood (LLH) and parameters for each data column

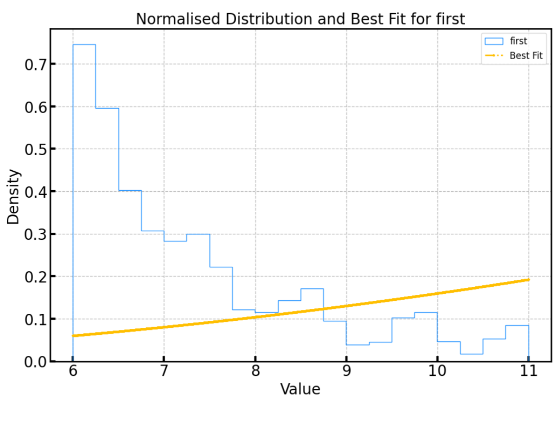

Figure 1: The distribution and fit of the first data.

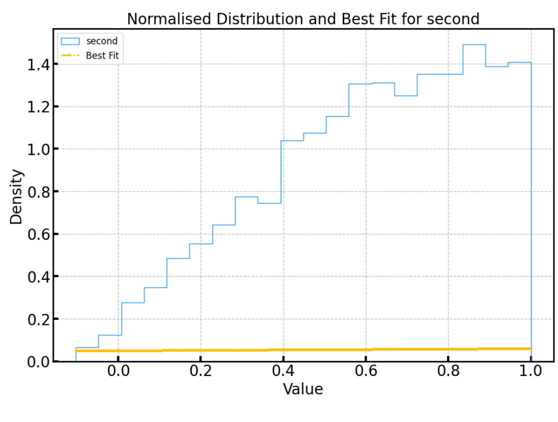

Figure 2: The distribution and fit of the second data.

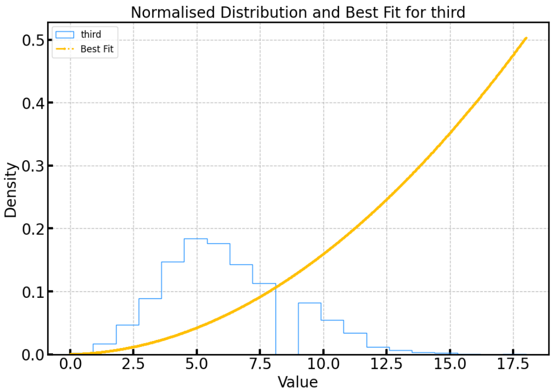

Figure 3: The distribution and fit of the third data.

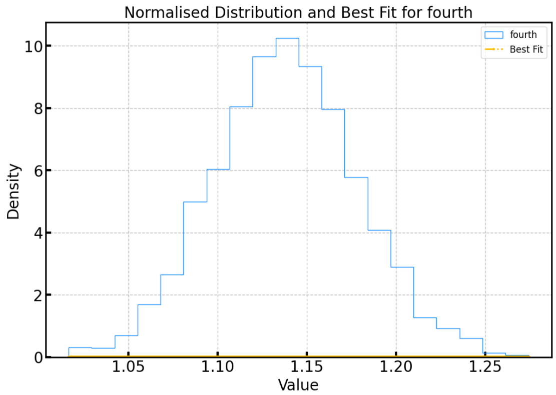

Figure 4: The distribution and fit of the fourth data.

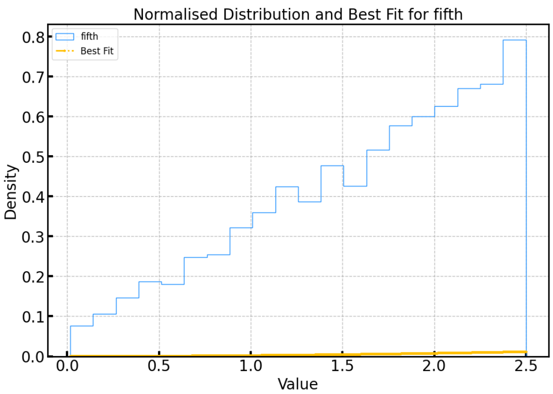

Figure 5: The distribution and fit of the fifth data.

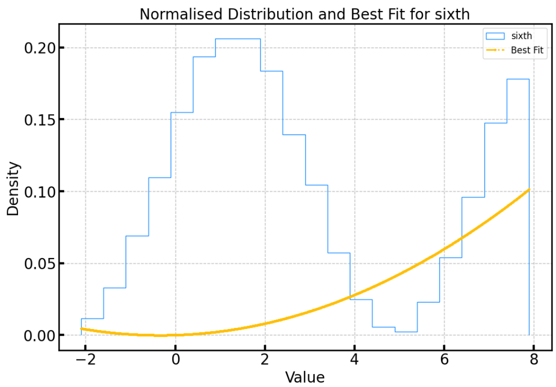

Figure 6: The distribution and fit of the sixth data.

PROBLEM 2

a

To assess the isotropy of the two data distributions, I perform a Kolmogorov-Smirnov (KS) test. First, I investigate the distribution of the original data in Fig7. Figure 8 displays the generated pseudo-data, based on the premise that angles are uniformly distributed across the board. While this theory holds some water for zenith angles, it seems less viable for azimuthal angles.

b

Figures 9 and 10 show the pseudo-data and the results of the KS test for the first (A) and second (B) alternative hypotheses, respectively. I can see that hypothesis A is better than the null hypothesis and hypothesis B, having the best p-values for the zenith angle(0.681) and the azimuthal angle(0.502).

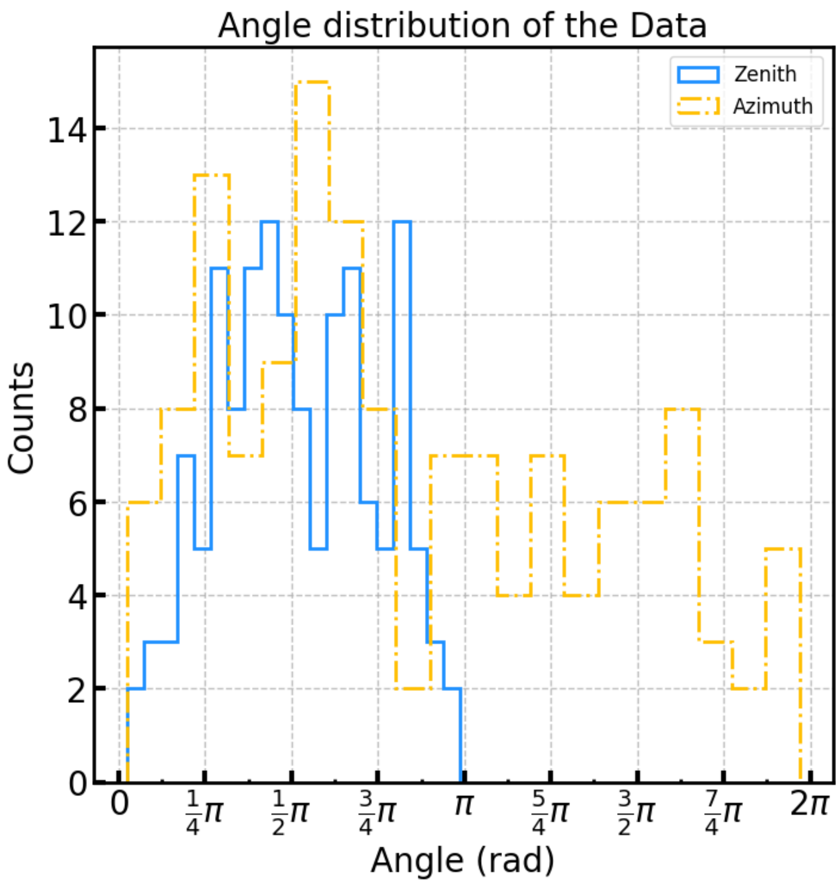

Figure 7: The distribution of the given data.

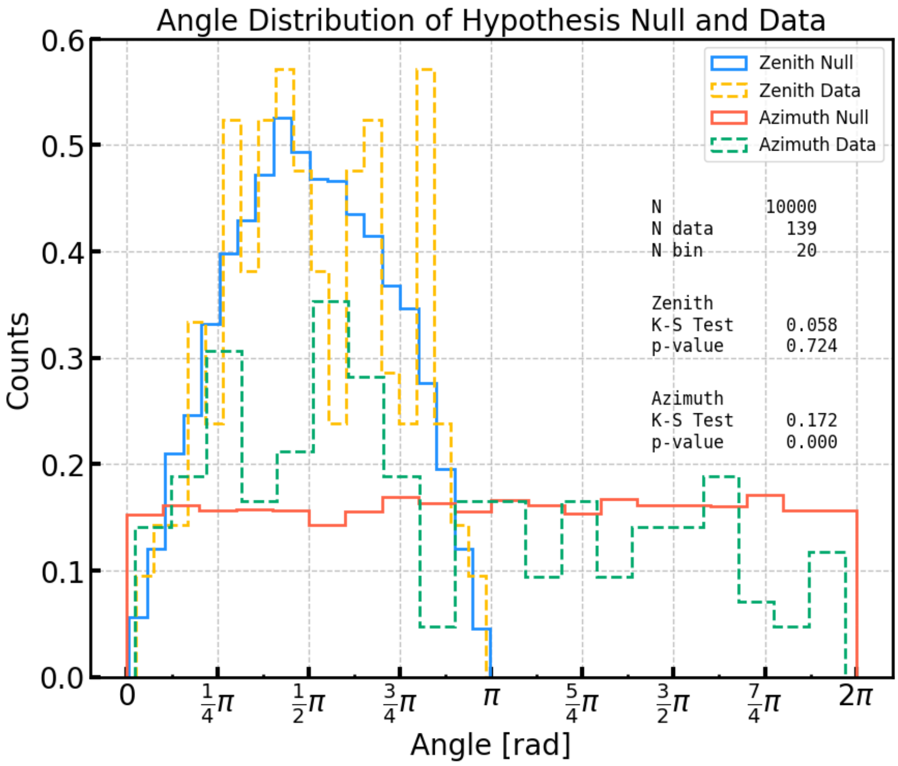

Figure 8: The distribution of the totally isotropic case.

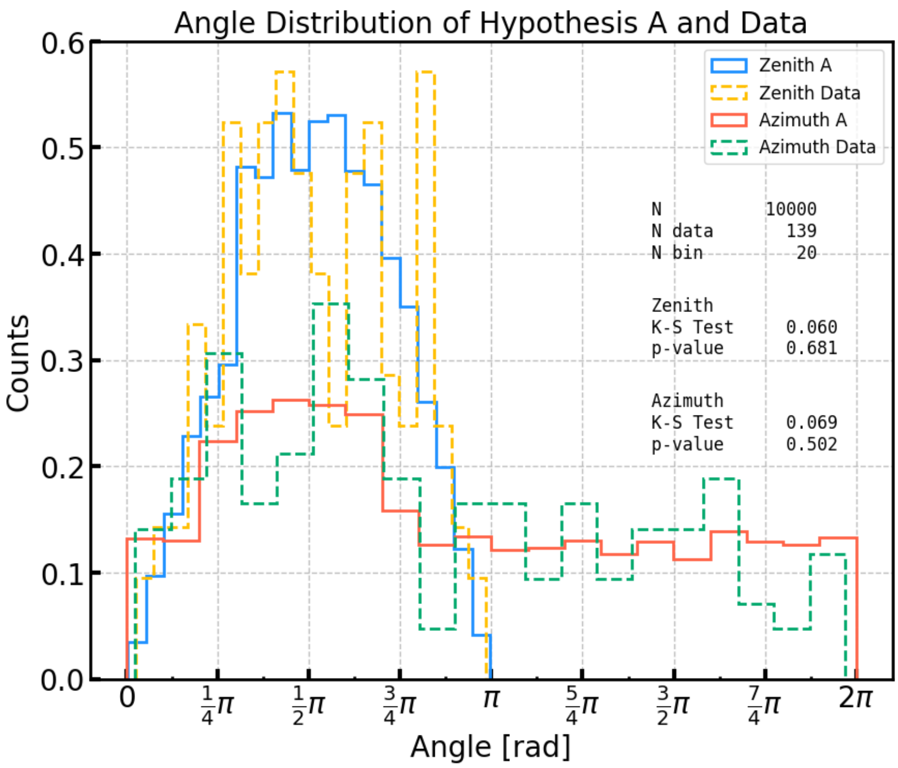

Figure 9: The distribution of the hypothesis A.

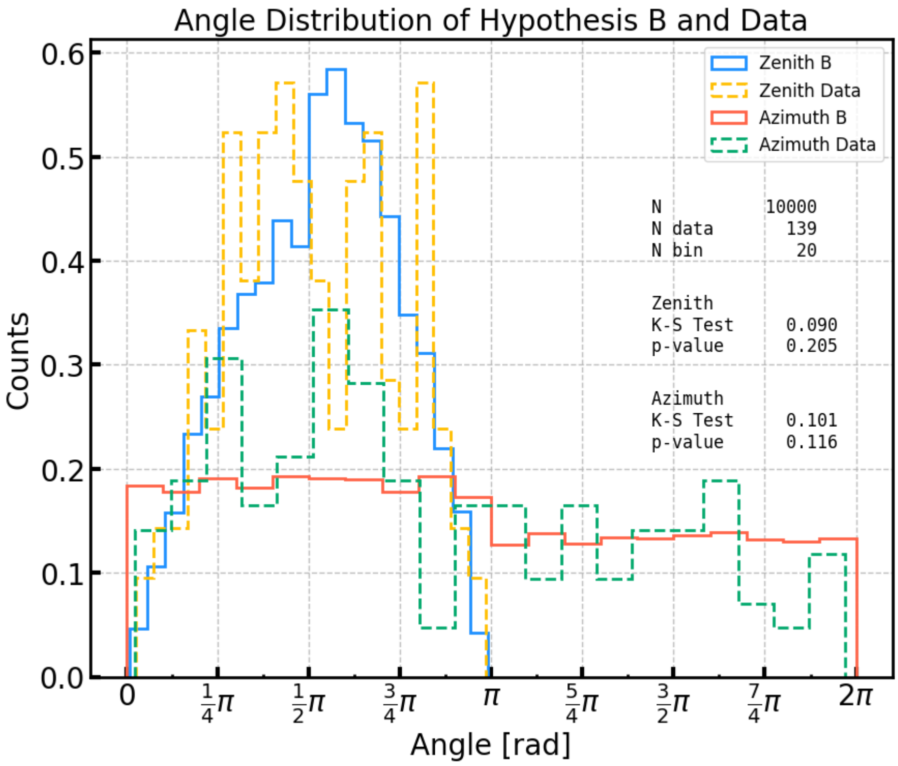

Figure 10: The distribution of the hypothesis B.

PROBLEM 3

$$\mathcal{L}(\theta_1, \theta_2, \theta_3) = 3\left(\cos(\theta_1)\cos(\theta_2) + \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\theta_3-\mu)^2}{2\sigma^2}}\right)\cos\left(\frac{\theta_1}{2}\right) + 3$$
$$\mu = 0.68$$
$$\sigma = \sqrt{0.04}$$

a

Utilising Python's *nestle* package, I employ the nested sampling algorithm to work with the given test statistics function. The parameters of the function θ_1, θ_2, and θ_3 have a finite range. As the function is not normalised, what I derive here isn't strictly a probability density function (pdf) or a Kernel Density Estimation (KDE), but rather an approximation of the posterior or likelihood. The optimal values for θ_1, θ_2, and θ_3 are illustrated in Figure 11 and Figure 12. These values tend to correspond to the local maxima of the pseudo-density estimation.

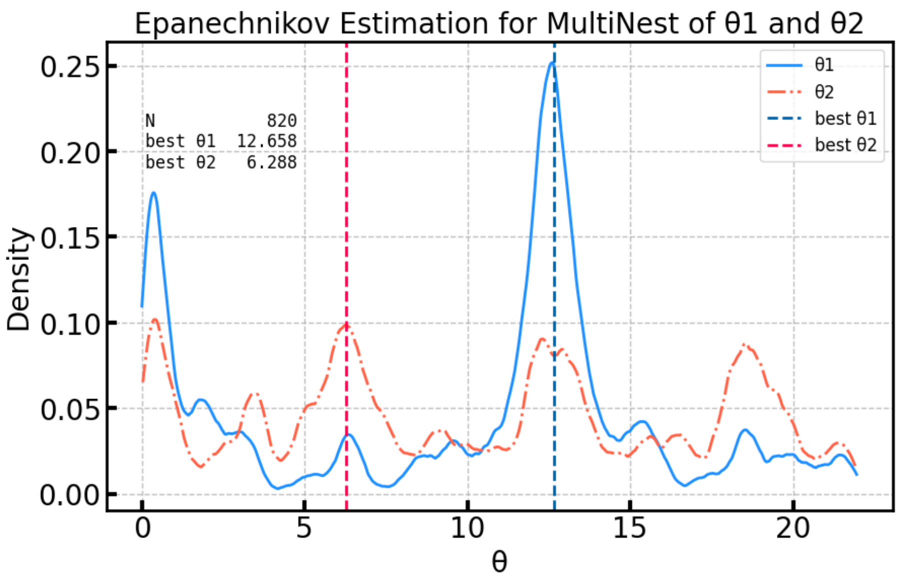

Figure 11: The pseudo density estimation of MultiNest of $\theta1$ and $\theta2$.

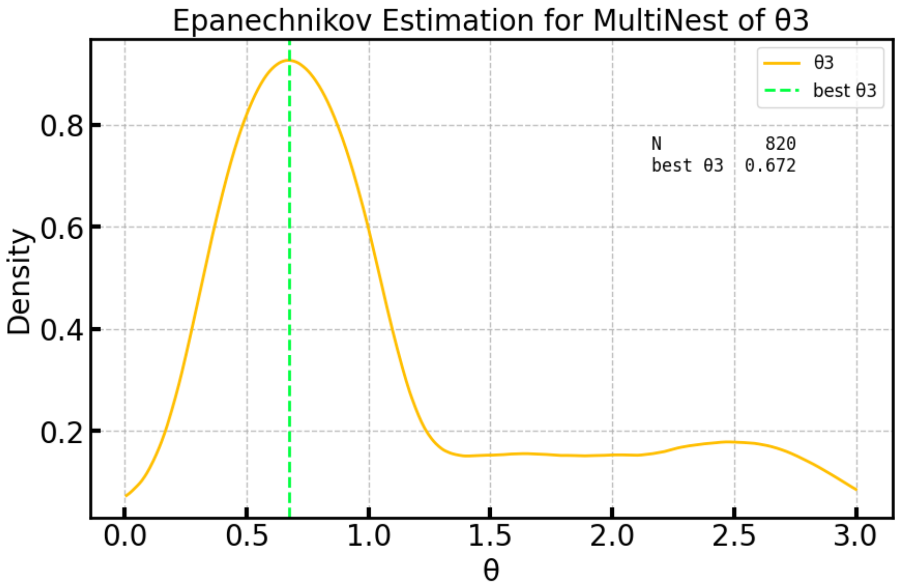

Figure 12: The pseudo density estimation of MultiNest of $\theta 3$.

b

The 2D KDE contours of the parameters θ_1, θ_2, and θ_3 are displayed in Figure 13 and Figure 14. Using this pseudo-posterior, I can ascertain the optimal parameter values on a 2D map.

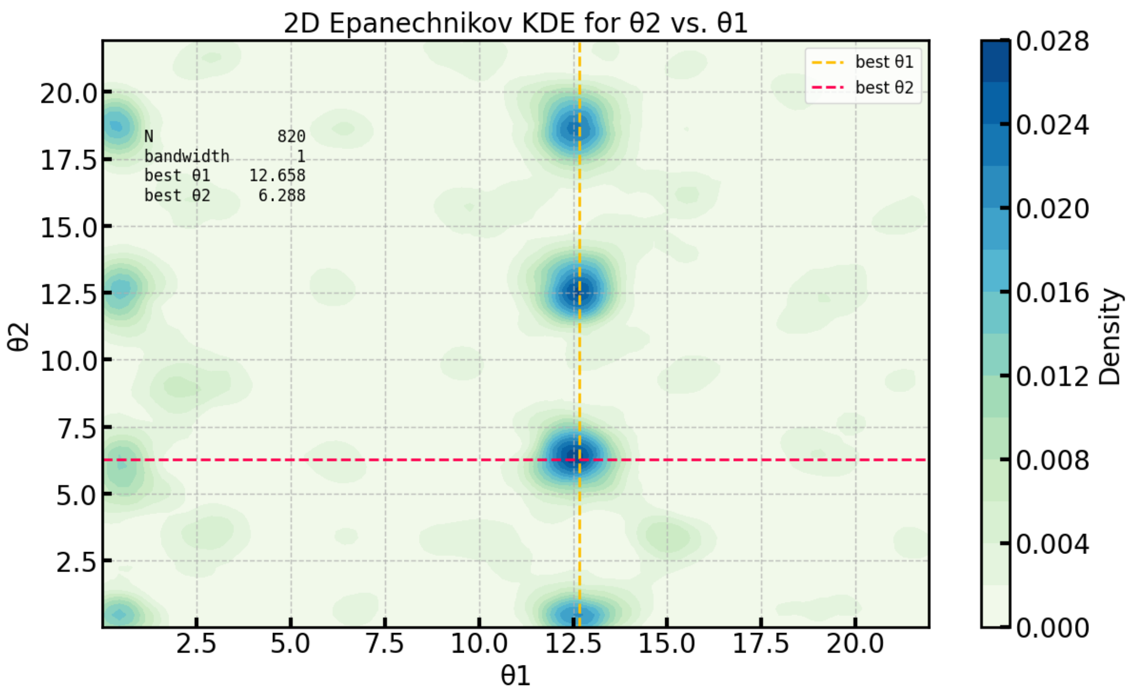

Figure 13: The 2D Epanechnikov KDE of MultiNest of $\theta 1$ and $\theta 2$.

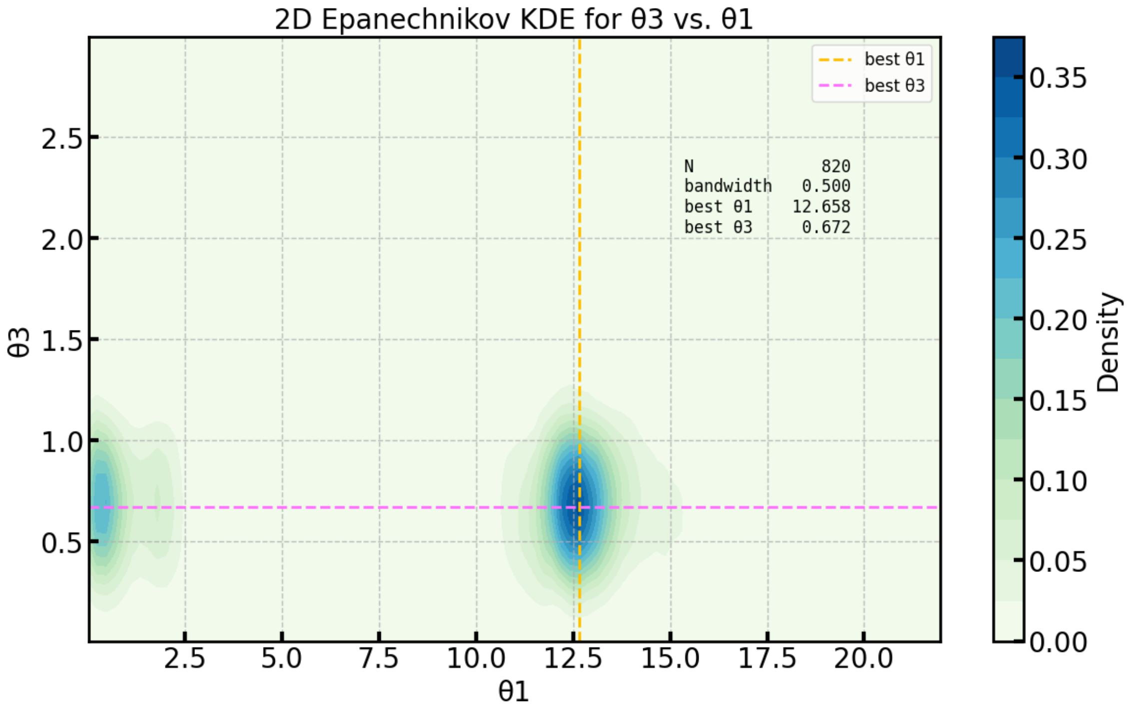

Figure 14: The 2D Epanechnikov KDE of MultiNest of $\theta 1$ and $\theta 2$.

c

The raster scan of the function's parameters is presented in Figure 15 and Figure 16, complete with detailed information. These values seem to correspond with the outcomes obtained from the MultiNest KDE. Nevertheless, considering MultiNest's random selection mechanism and the finite bin size of the raster scan, exact matches in values may not be achieved.

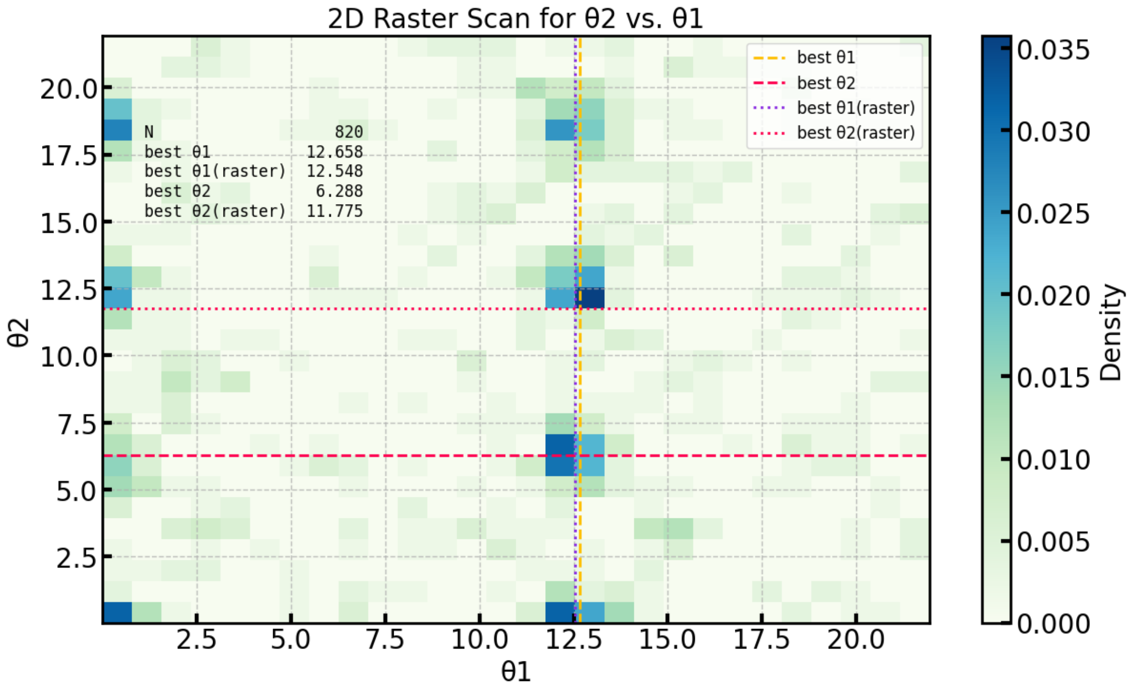

Figure 15: The 2D raster scan of MultiNest of $\theta1$ and $\theta2$.

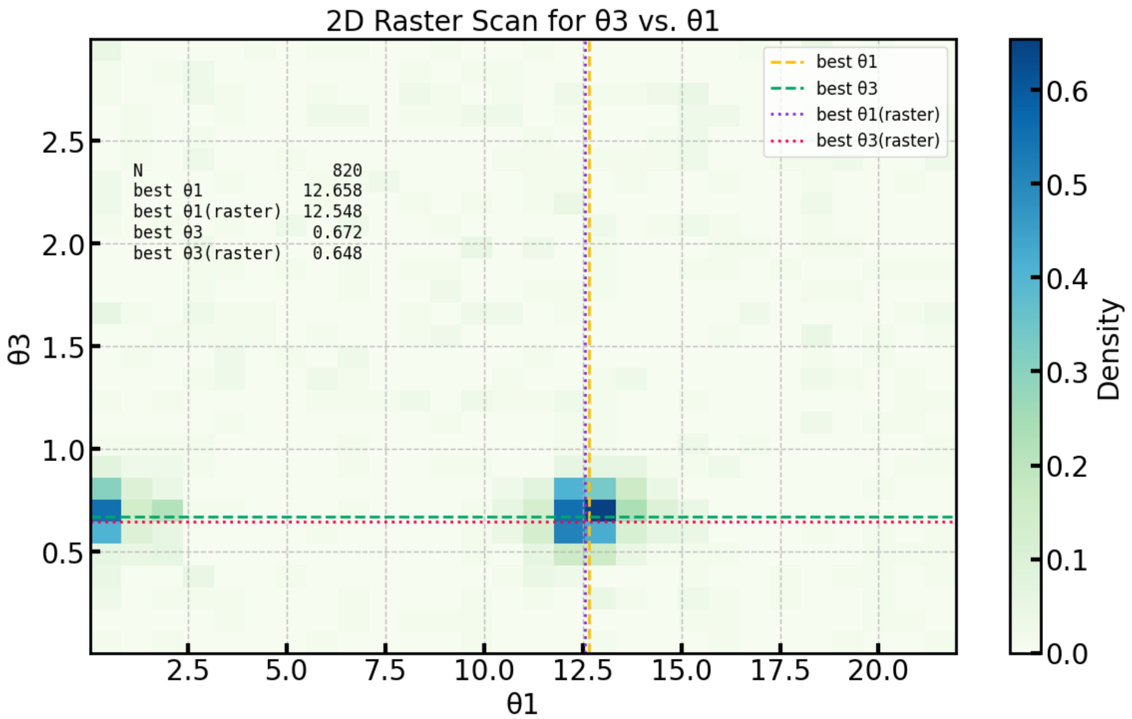

Figure 16: The 2D raster scan of MultiNest of $\theta1$ and $\theta3$.

PROBLEM 4

I utilised five classifiers for this problem: Logistic Regression, Random Forest, Adaptive Boost, Gradient Boosting, and XGB. It seems Gradient Boosting gives the best result so I enclose the result achieved by Gradient Boosting.

a

The classified test data are shown in Figure17, 18, 19, 20, and 21 with the details. The confusion matrix, accuracy, precision and the feature ranking are also shown in the figures.

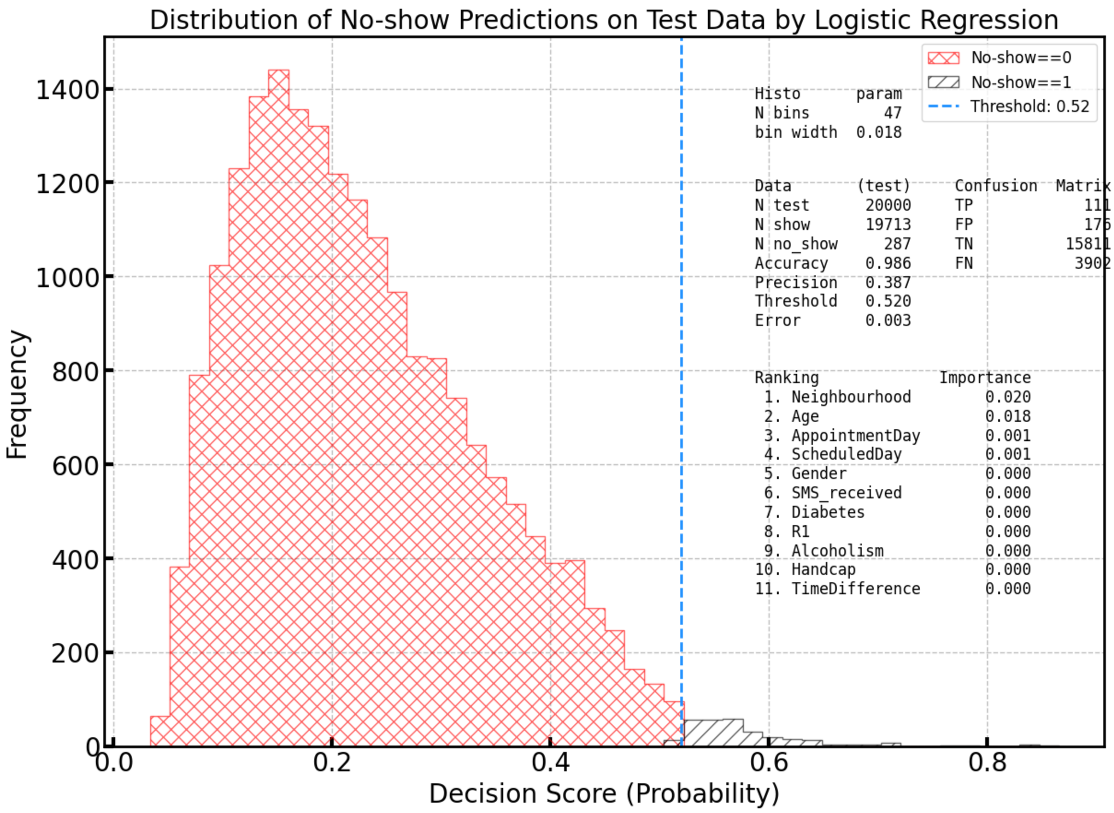

Figure 17: Classification of the test data by Logistic Regression.

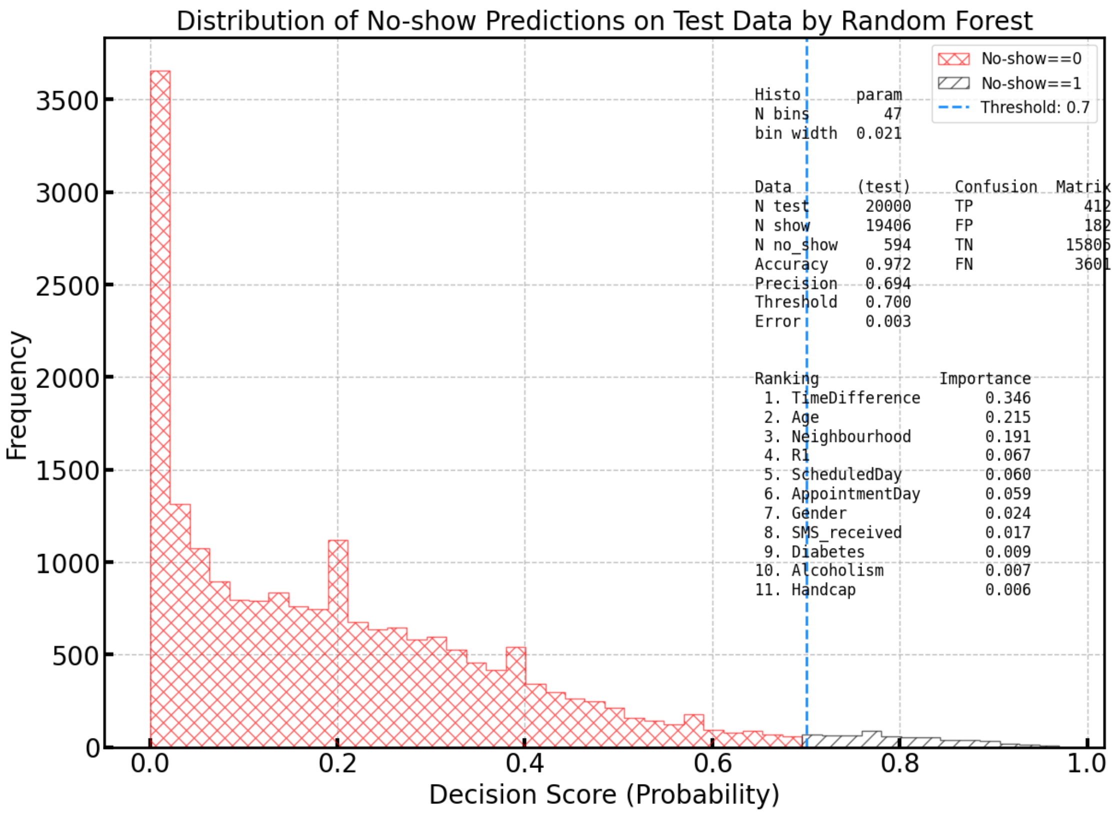

Figure 18: Classification of the test data by Random Forest.

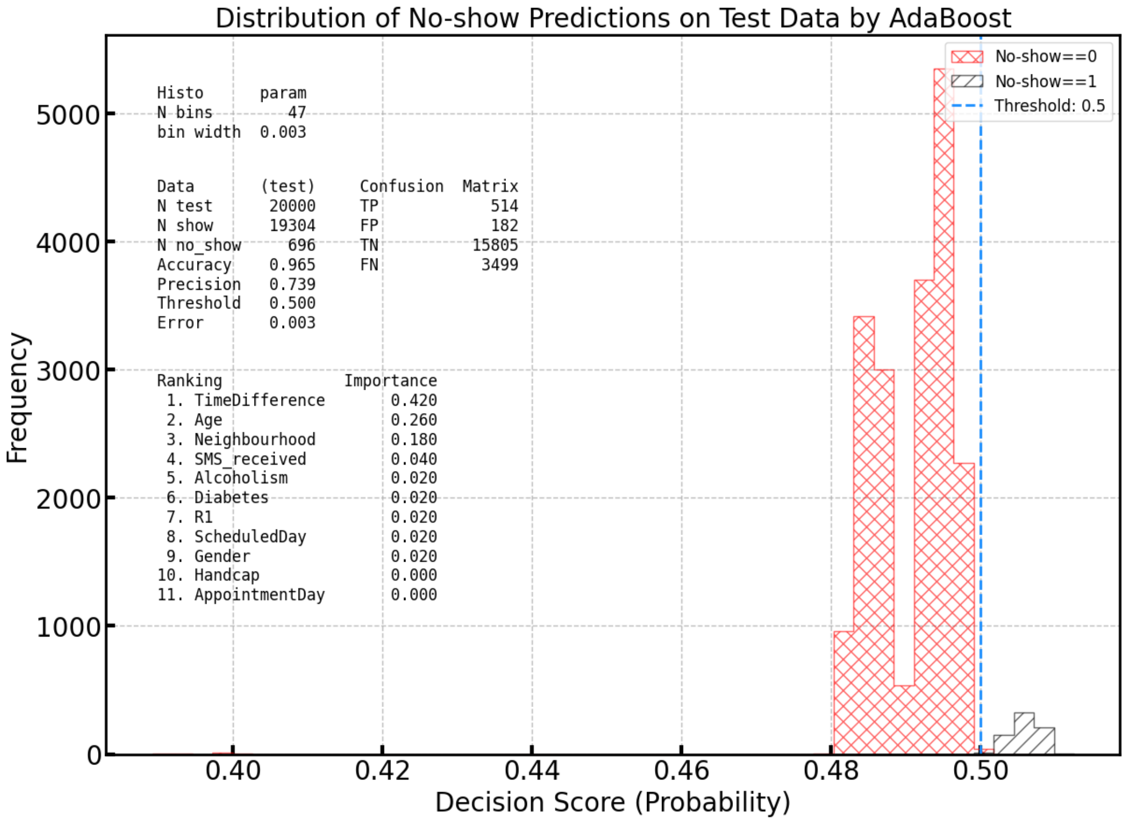

Figure 19: Classification of the test data by Adaptive Boosting.

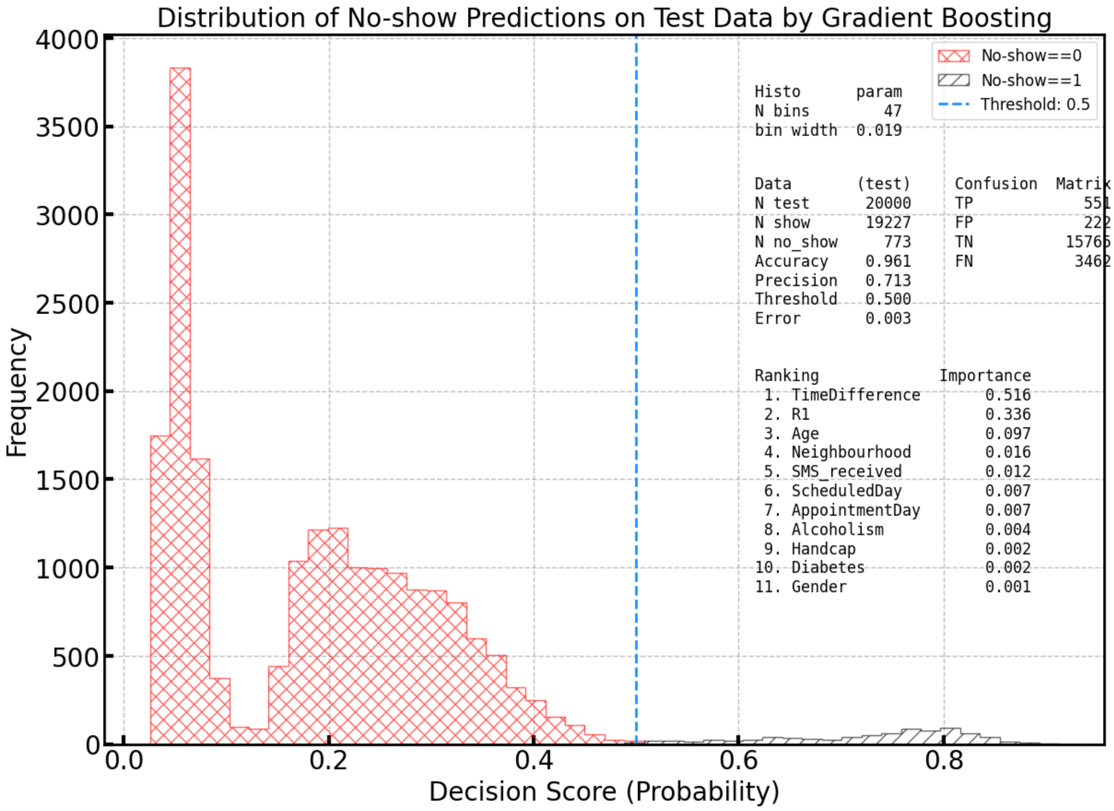

Figure 20: Classification of the test data by Gradient Boosting.

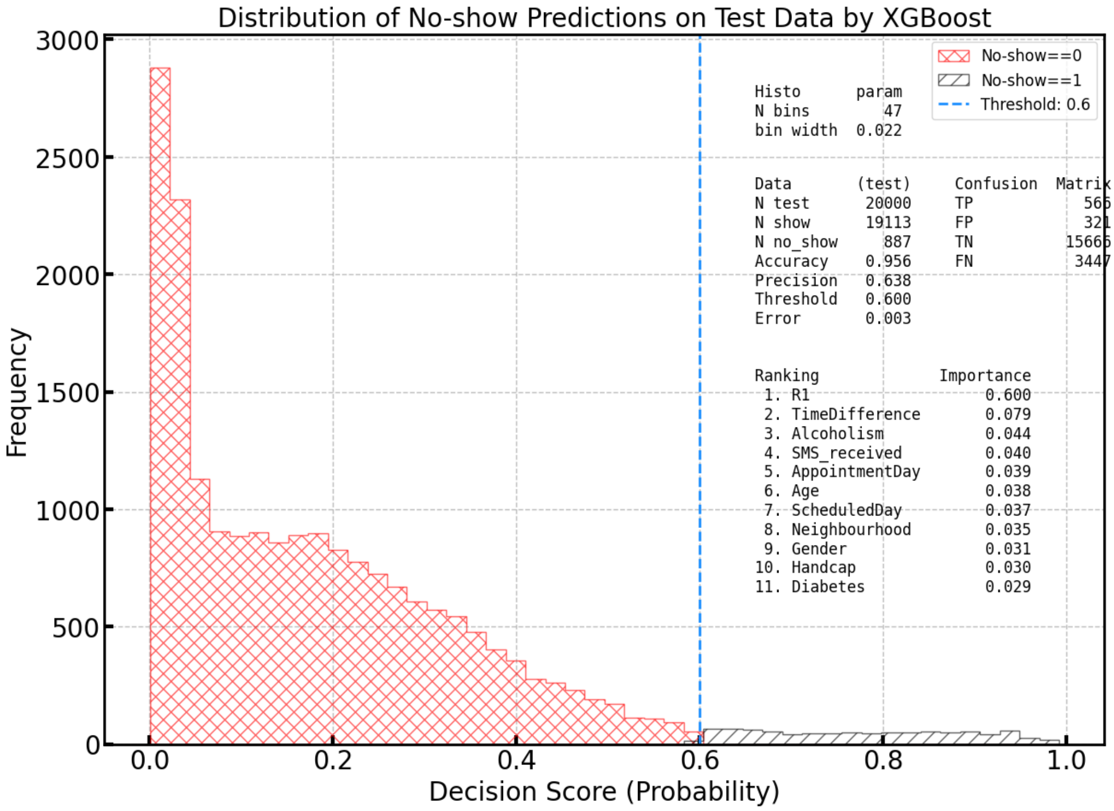

Figure 21: Classification of the test data by XGB.

b

The classified blind data are shown in Figure22, 23, 24, 25, and 26 with the details. The confusion matrix, accuracy, precision and feature ranking are also shown in the figures. The accuracies of the classifiers are shown in TableII.

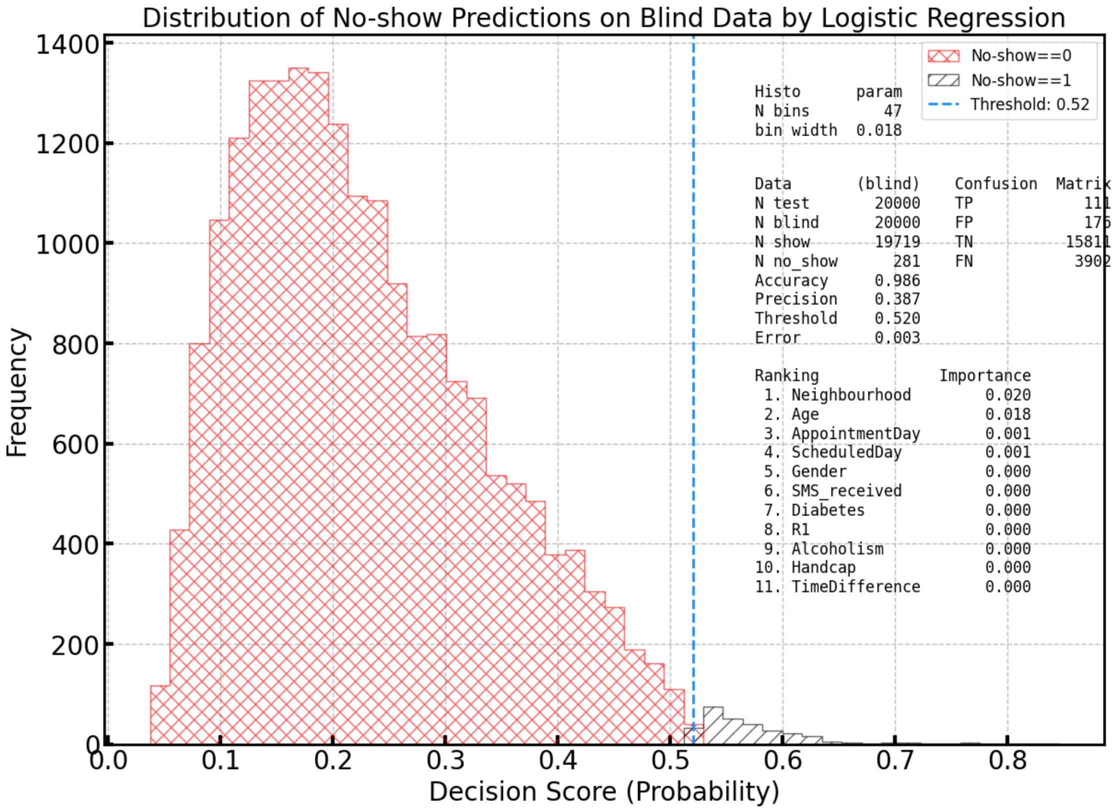

Figure 22: Classification of the blind data by Logistic Regression.

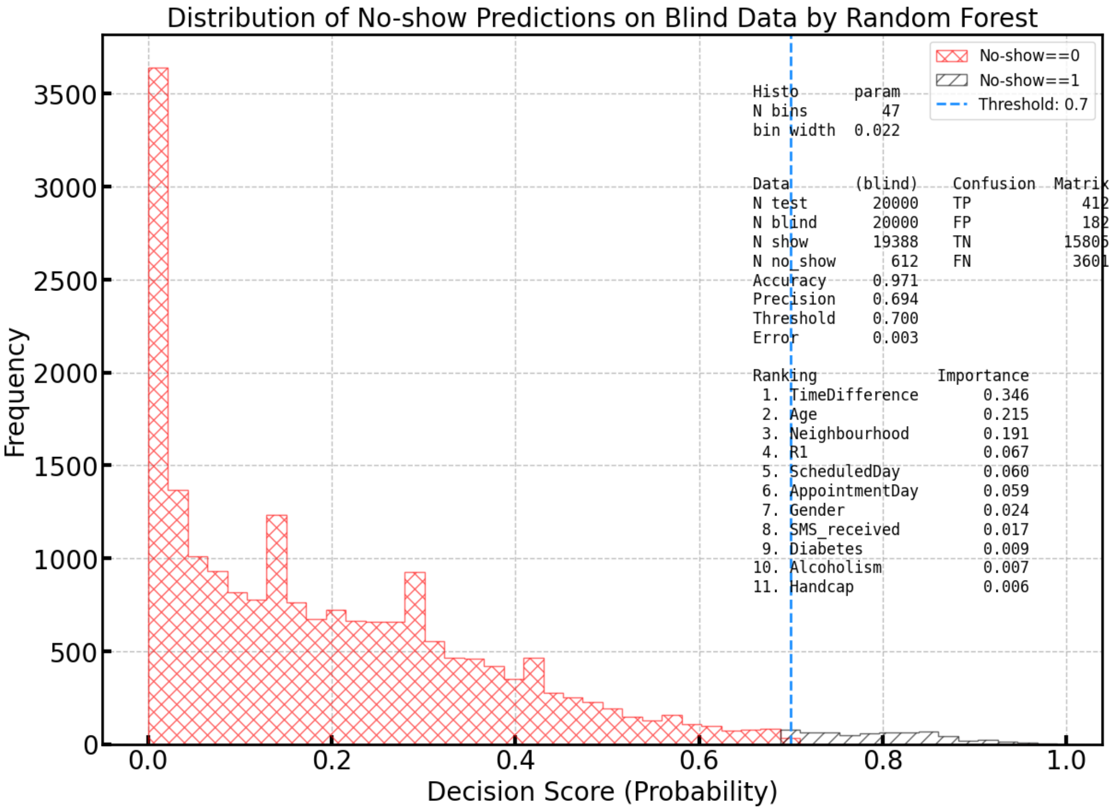

Figure 23: Classification of the blind data by Random Forest.

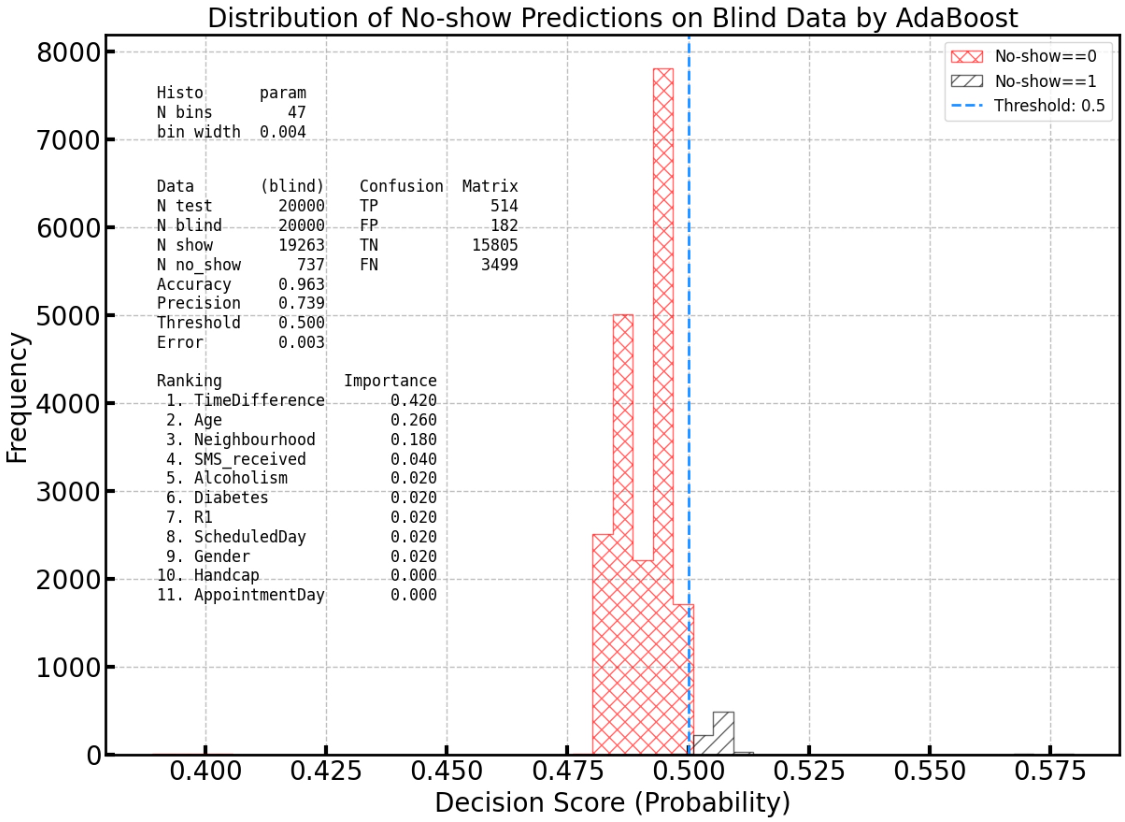

Figure 24: Classification of the blind data by Adaptive Boosting.

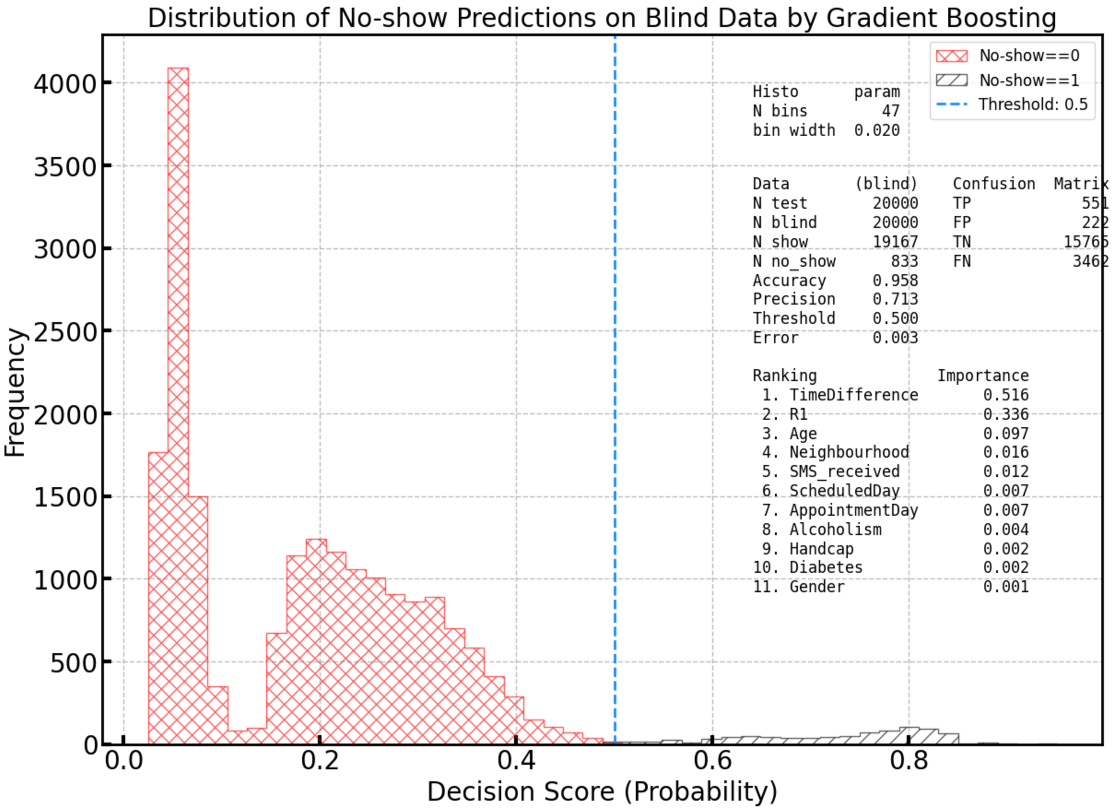

Figure 25: Classification of the blind data by Gradient Boosting.

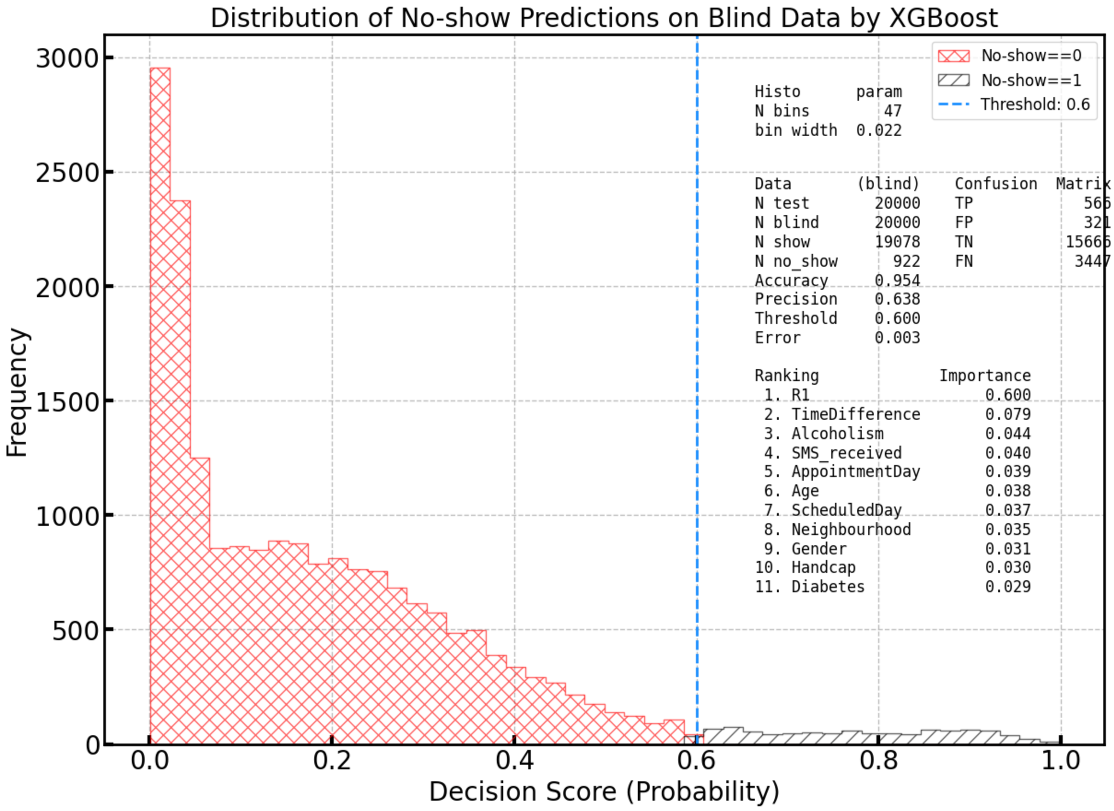

Figure 26: Classification of the blind data by XGB.

Classifier	**Accuaracy**
Logistic Regression	0.986
Random Forest	0.971
Adaptive Boost	0.963
Gradient Boosting	0.958
XGB	0.954

Table II: The accuracy of the classification done by each classifier

PROBLEM 5

a

Figures 27 and 28 display both the linear and cubic splines, along with the estimated temperature values at 203.570 sols. The temperature estimation using the linear spline yields $-115.913C°$, whereas the cubic spline yields $-115.325C°$.

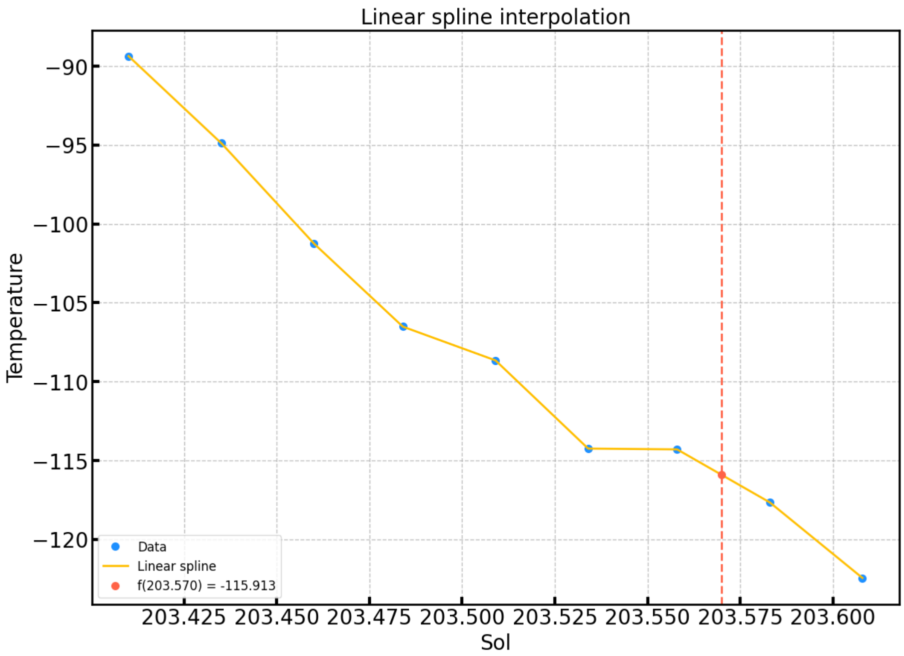

Figure 27: Linear spline of the data.

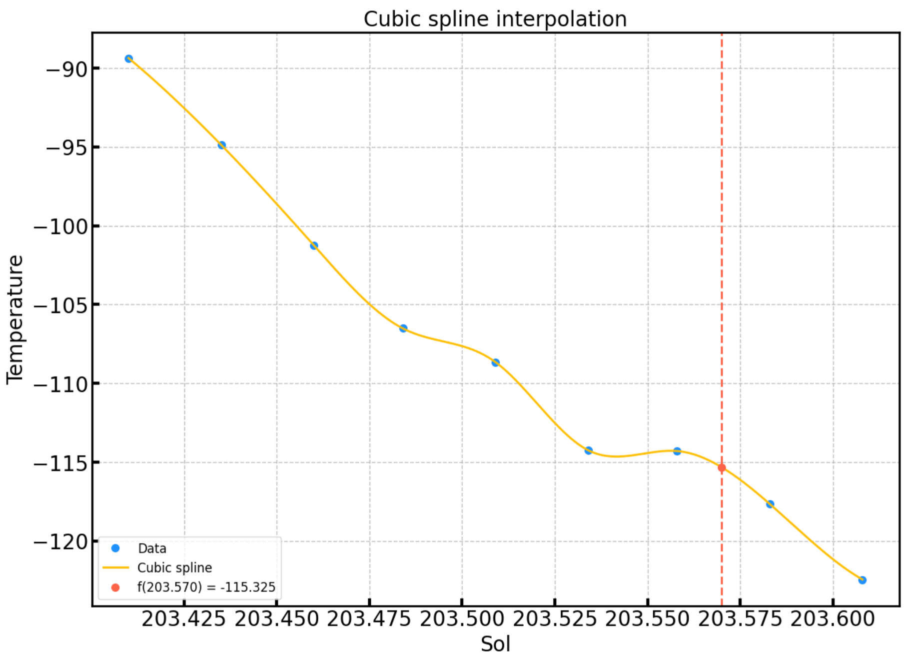

Figure 28: Cubic spline of the data.

b

The interpolation scatter plot is shown in Figure29. There is an increasing zone for the cubic spline between 203.542 and 203.556 sol, so it may require further investigation.

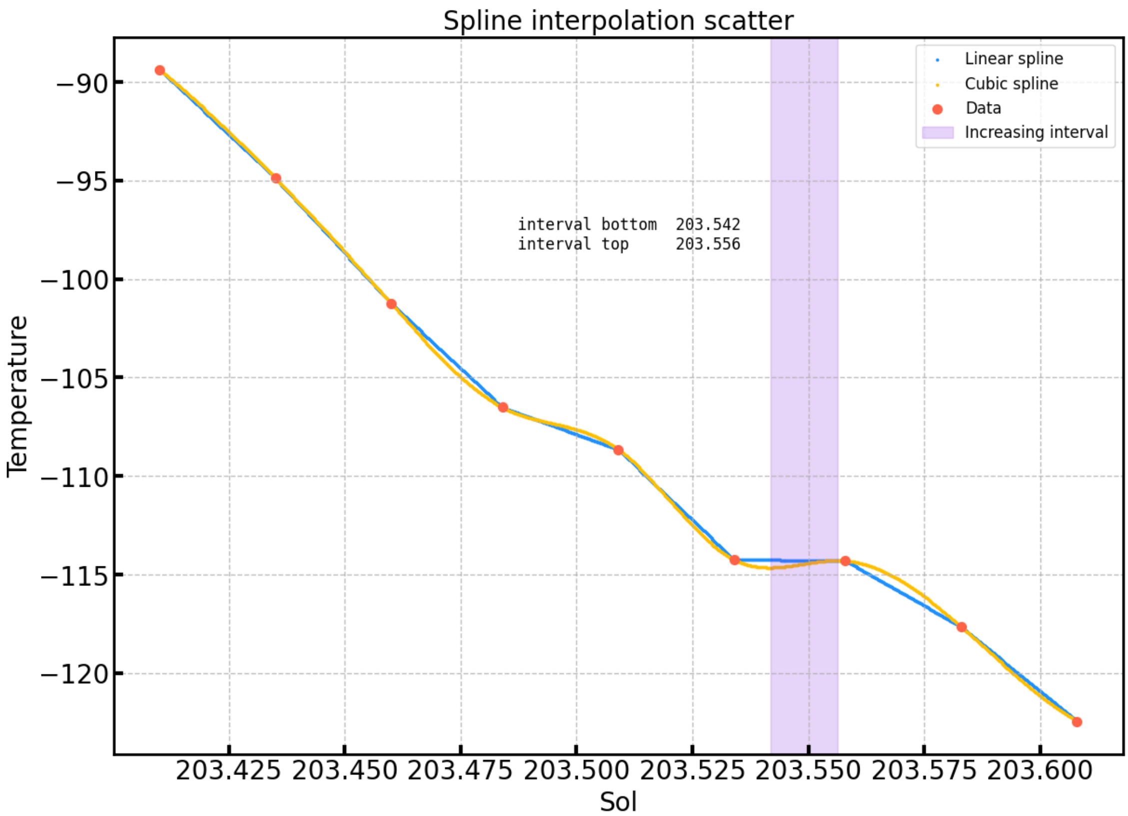

Figure 29: Scatter plot of the interpolations by two splines.

c

The temperature change rate, computed from the two splines, is presented in Table III along with the threshold condition. It seems that both splines indicate the temperature change exceeds the hardware specifications, suggesting a need for improvement.

Spline Type	**Max Rate of Change (°C per sol)**	**Sustainable**
Linear	254.80	No
Cubic	265.64	No

Table III: Temperature Change Sustainability with Threshold of 227.50 °C per sol