



**Train a Machine to Find the Rare Decay**  
 **$B^0 \rightarrow K^*(892)^0 \mu^+ \mu^-$  at the Belle Experiment**  
**DESY Summer School 2017**

Margarete Kattau, Ludwig-Maximilians-University Munich, Germany

September 6, 2017

**Abstract**

This report sums up my work as a participant at the DESY Summer School 2017. The project has been executed at the Belle II group and involved the analysis of Monte Carlo data of the Belle I experiment. The focus was placed on the rare decay  $B \rightarrow K^{(*)} \mu^+ \mu^-$ . The classification of signal and background events has been performed using machine learning methods. Different methods have been compared and evaluated. Concluding, the branching ratio for the  $B \rightarrow K^{(*)} J/\psi$  channel has been calculated.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Analysis</b>	<b>3</b>
2.1	Signal Cuts . . . . .	3
2.2	Background Sources . . . . .	3
2.3	Selection of Variables . . . . .	5
2.4	Comparison of Classifiers . . . . .	5
2.5	Calculation of the Branching Ratio . . . . .	12
<b>3</b>	<b>Summary</b>	<b>13</b>
<b>A</b>	<b>Overview of Training Variables</b>	<b>15</b>

# 1 Introduction

The search for new physics beyond the standard model is an essential part of modern physics. Although the standard model is a well-established theory, there are some issues which cannot be solved by it alone. One of those is the so-called baryon asymmetry that describes the imbalance in baryonic and antibaryonic matter which we observe in the universe.

The Belle experiment at KEK in Tsukuba in Japan searches for signs of new physics by colliding electrons and positrons at energies near the  $\Upsilon(4S)$  resonance which leads to the pairwise production of B mesons. In order to find hints on new physics the analysis of rare decays is essential because in this case new physics can occur in the same order of magnitude as standard model processes. The asymmetric collider KEKB has a circumference of 3 km. The data was taken between 1999 and 2010. In order to reach a higher luminosity Belle is currently upgraded to Belle II which is expected to start first physical runs in 2018. [1]

Within the scope of this Summer School project simulated Monte Carlo events were used to analyse the rare decay  $B \rightarrow K^{(*)}\mu^+\mu^-$ .

## 2 Analysis

The analysis of the simulated data is performed in Python using *Jupyter Notebook*. To obtain a high signal yield and an effective background suppression the package *scikitlearn* is used. It provides a selection of various classifiers for machine learning.

### 2.1 Signal Cuts

The applied cuts are adapted from [1]. They are necessary to suppress background events originating from charmonium decays like  $B \rightarrow K^{(*)}J/\psi$  and  $B \rightarrow K^{(*)}\psi(2S)$ , where the  $c\bar{c}$  state decays into two leptons [1]. Within the scope of this project only the muon final states are considered. The following cuts are applied:

$$-0.15 \frac{GeV}{c^2} < M_{\mu\mu} - M_{J/\psi} < 0.08 \frac{GeV}{c^2}, \quad (1)$$

$$-0.10 \frac{GeV}{c^2} < M_{\mu\mu} - M_{\psi(2S)} < 0.08 \frac{GeV}{c^2}. \quad (2)$$

### 2.2 Background Sources

For the analysis of  $B \rightarrow K^{(*)}\mu^+\mu^-$  several background sources have to be considered [1]:

- Continuum: Events arising from this background source originate from  $e^+e^-$  annihilation. The generated particles are the light quark pairs  $u\bar{u}$ ,  $d\bar{d}$ ,  $s\bar{s}$  and  $c\bar{c}$ . In this analysis several training variables are used to suppress those background events.

- **Combinatorial:** These background events originate from wrong combinations of tracks in B decays and are the dominant source of background.
- **Peaking:** Processes imitating the signal shape in  $M_{bc}$  are called peaking background processes. Among them are the channels for the irreducible background sources  $B \rightarrow K^{(*)}J/\psi$  and  $B \rightarrow K^{(*)}\psi(2S)$  that have been described in section 2.1. Moreover, pions from the decay  $B \rightarrow K^{(*)}\pi\pi$  can be mistakenly treated as muons.
- **Cross-feed:** The reason for this source are candidates that have been assigned to the wrong decay channel due to missing decay products or misreconstruction of one of their children.

Further background events arise from random combinations in the charged ( $B^+B^-$ ) and mixed ( $B^0\bar{B}^0$ ) generic Monte Carlo. Figure 1 shows the distribution of the so called Q value which describes the energy difference between the mother and decay particles.

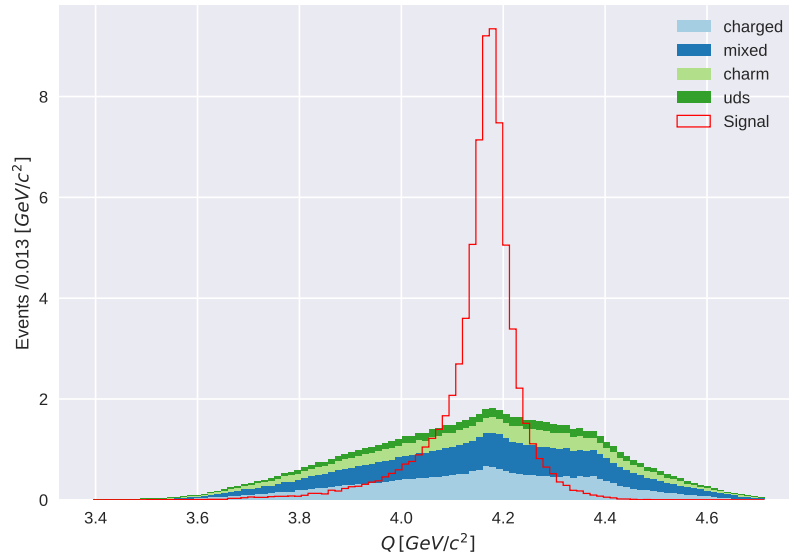


Figure 1: Distribution of signal and background events for the Q value from the channels **charged**, **mixed**, **charm** and **uds**. The **charged** and **mixed** channels contain events arising from random combinations of  $B^+B^-$  and  $B^0\bar{B}^0$  decays in the generic Monte Carlo. The other two describe the continuum background events  $e^+e^- \rightarrow c\bar{c}$  and  $e^+e^- \rightarrow u\bar{u}, d\bar{d}, s\bar{s}$ .

## 2.3 Selection of Variables

In order to find variables suitable for the separation of signal and background, both distributions are plotted in the same histogram for every variable. Figure 2 shows the distribution for the variable  $\Delta E = E_B - E_{Beam}$ , which is useful for discriminating signal against background.

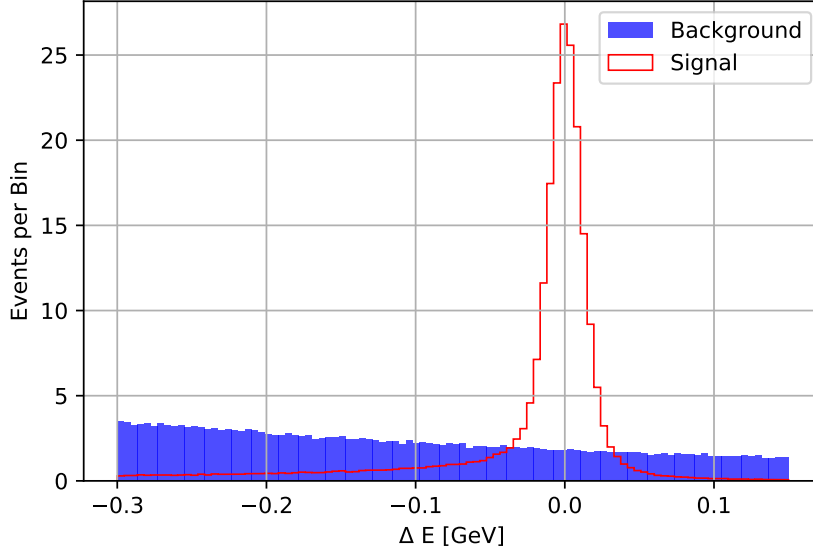


Figure 2: Comparison between signal and background events for the variable  $\Delta E$ . It describes the difference between the energy of the B mesons and the energy of the beam [1].

In addition, correlation matrices are created to check if the selected variables are correlated to the beam constrained mass

$$M_{bc} = \sqrt{E_{Beam}^2 - |\vec{p}_B|^2}, \quad (3)$$

where  $\vec{p}_B$  is the momentum of the reconstructed candidate.  $M_{bc}$  features a signal distribution suitable to discriminate against background. Therefore, the correlation between  $M_{bc}$  and the selected variables should be as small as possible. [1]

In order to find a good variable set different combinations of variables are tested. Table 2 in the appendix gives an overview of the training variables.

## 2.4 Comparison of Classifiers

*Scikitlearn* offers a wide range of different classifiers for machine learning. In the described project they are used to distinguish signal from background. To avoid biasing, the data set on which the classifier is trained on should always differ from the one that is supposed to be classified. At first, different classifiers are chosen and compared:

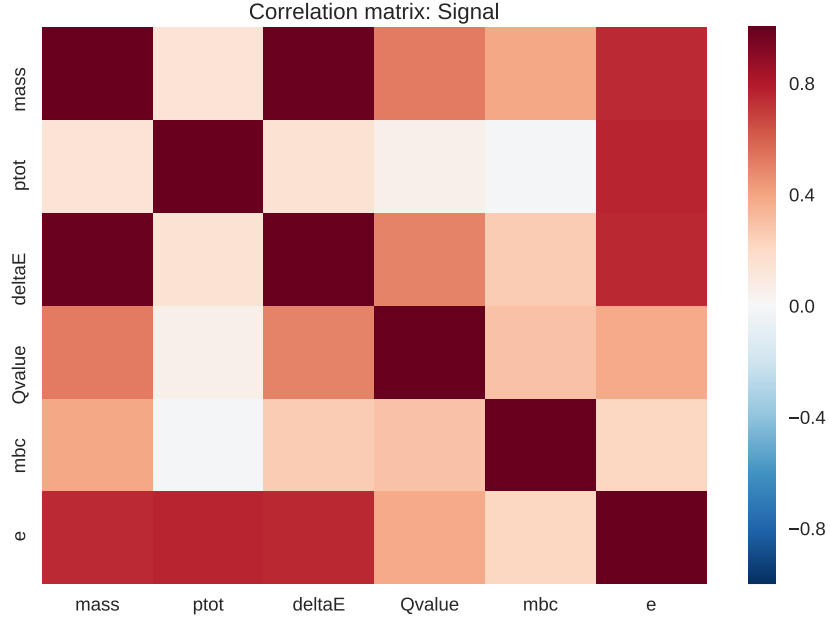


Figure 3: Correlation matrix of the signal. The plot shows the correlation between the different variables from an example training set. The legend shows the degree of correlation. For our analysis a correlation between the beam constrained mass  $M_{bc}$  and the other variables used for the training should be avoided.

- Boosted Decision Tree Classifier (bdt)
- Neural Network Classifier (mlp)
- Nearest Neighbours Classifier (nn)
- Random Forest Classifier (rf)
- Ada Boost Classifier (ab)
- Gaussian Naive Bayes Classifier (gnb)
- Quadratic Discriminant Analysis Classifier (qda)

For every classifier the Receiver Operating Characteristics (=ROC) Curve is created. It is a helpful tool to evaluate the effectiveness of a classifier and shows the correlation between the True Positive Rate and the False Positive Rate. Table 1 describes the different terms.

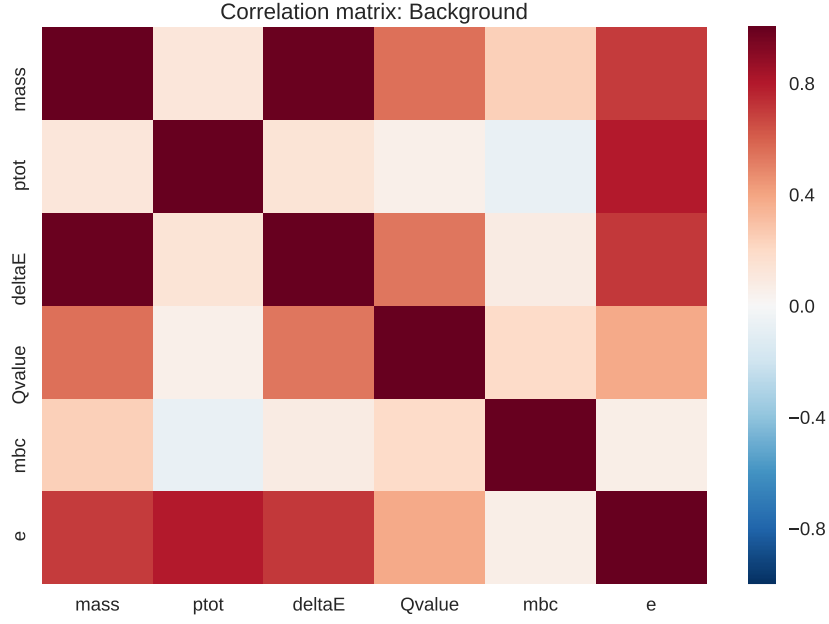


Figure 4: Correlation matrix of the background for the same set of variables as figure 3.

Actual Class	Classification Result	
	True	False
True	True Positive	False Negative
False	False Positive	True Negative

Table 1: Different terms necessary to evaluate a classifier. The column "Actual Class" describes the real class of an object whereas the column "Classification Result" describes the output of the classification process. The table is derived from [2].

The True Positive Rate describes the amount of correctly classified signal events compared to all positively classified objects [2]:

$$TPR = \frac{\#True\ Positive}{\#True\ Positive + \#False\ Positive}. \quad (4)$$

The False Positive Rate is defined accordingly to that. Figure 5 shows the ROC curve of the Boosted Decision Tree classifier whereas Figure 6 displays the ROC curves of all classifiers named above.

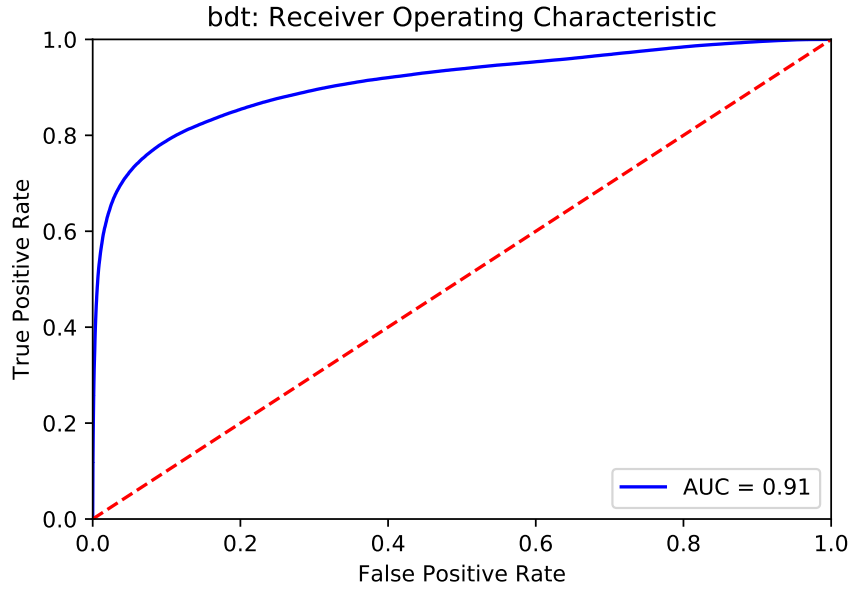


Figure 5: Receiver Operating Characteristics plot for the Boosted Decision Tree Classifier. The essential value in this figure is the area under curve (=AUC).

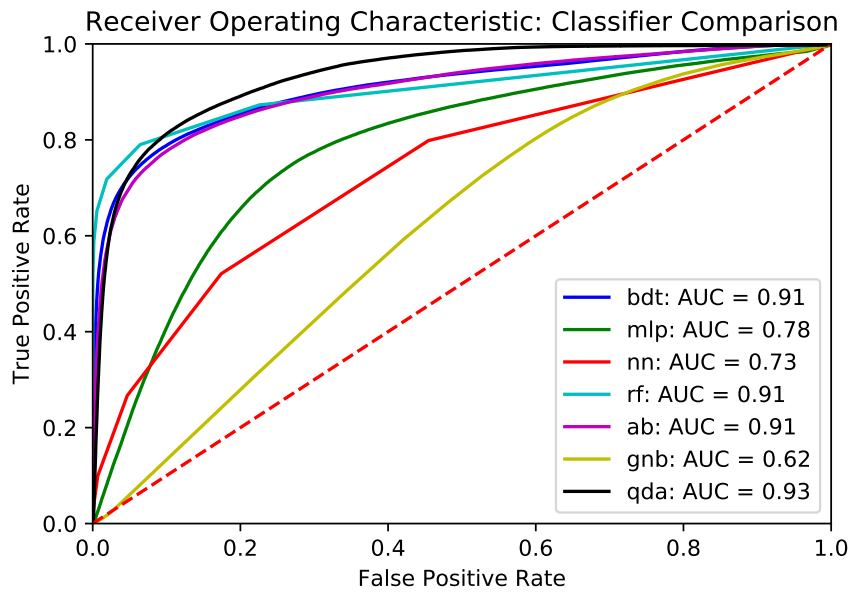


Figure 6: Receiver Operating Characteristics plot for all classifiers described above. One can see that the Boosted Decision Tree and the Random Forest Classifier give a better value than, for example, the Nearest Neighbours Classifier.



Due to their high values for the area under the ROC curve (=AUC) one can conclude that the Boosted Decision Tree and the Random Forest Method give the best results. Therefore they are chosen as the preferred classifiers for the subsequent analysis. The next step is to compare different cuts on the classifier prediction which tells us the probability of the prediction to be true.

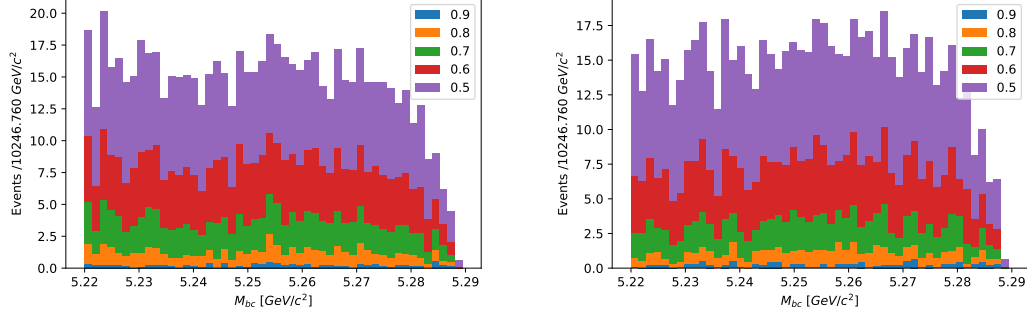


Figure 7: Classifier comparison for the background events. The histograms show the number of events remaining after a cut on the classifier prediction.

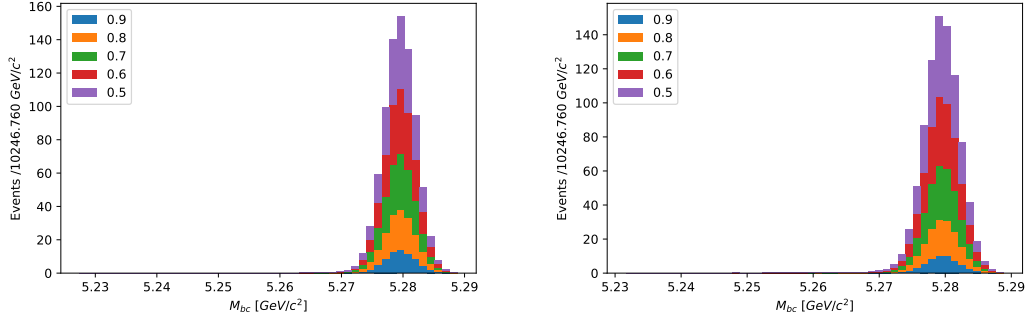


Figure 8: Classifier comparison for the signal events. The histograms show the number of events remaining after a cut on the classifier prediction.

Figures 7 and 8 show the corresponding plots. One can see the amount of remaining events after the cut. The aim is to obtain a small number of background events compared to the number of signal events received using the same cut. Moreover, the Figure of Merit (=FOM) is computed and plotted. It helps to find the best cut on the predicted probability for each classifier and is defined as:

$$FOM = \frac{n_{sig}}{\sqrt{n_{sig} + n_{bkg}}}, \quad (5)$$

where  $n_{sig}$  and  $n_{bkg}$  are the expected numbers of signal and background events in the region  $M_{bc} > 5.27 \frac{GeV}{c^2}$ . By extracting the maximum value of the Figure of Merit one

obtains the optimal cut that returns the highest statistical sensitivity for signal over background. Figure 9 shows the Figure of Merit as a function of the classifier prediction for the Boosted Decision Tree and the Random Forest Classifier.

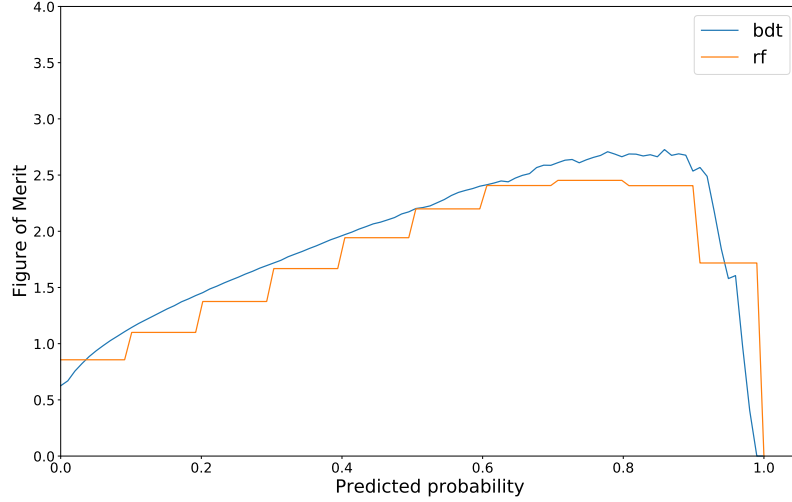


Figure 9: Figure of Merit as a function of the predicted probability for the Boosted Decision Tree and the Random Forest Classifier. The values for the optimal cuts are:  $c_{bdt} = 0.86$ ,  $c_{rf} = 0.80$ .

After finding the values for the optimal cuts they can be used to discriminate signal against background and obtain the signal yields. Figure 10 shows the correspondent distribution for both of the classifiers.

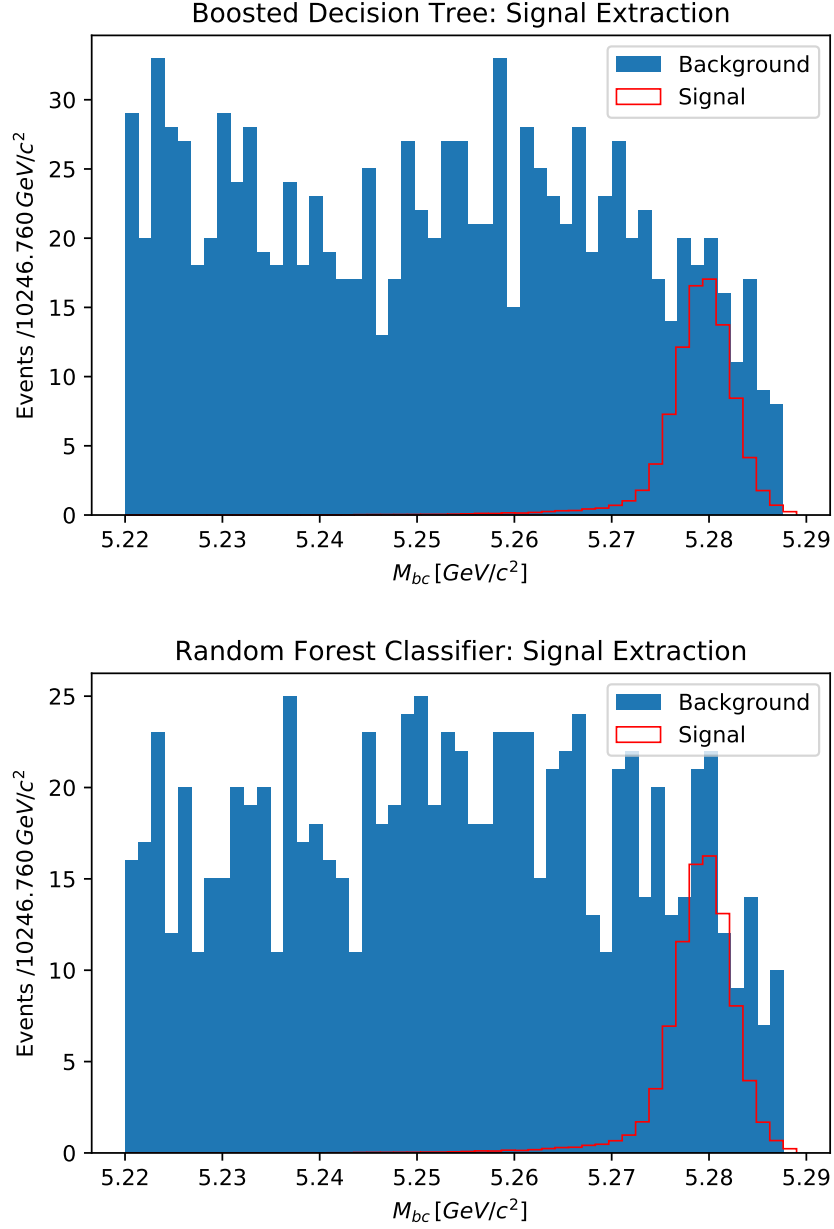


Figure 10: Signal yields for  $M_{bc}$  obtained with cuts on the classifier predictions for both the Boosted Decision Tree and the Random Forest Classifier. The signal is scaled down by the ratio of the expected number of signal events and the length of the whole signal data set. The cut values derived from the Figure of Merit are:  $c_{bdt} = 0.86$ ,  $c_{rf} = 0.80$ .

## 2.5 Calculation of the Branching Ratio

For the previous analysis areas were excluded due to a high number of background events which are described in section 2.1. In order to find the number of signal events in the veto area given by equation (1), the cut  $9 \text{ GeV}^2/c^4 < q^2 < 10 \text{ GeV}^2/c^4$  is applied. Because of its high efficiency the Boosted Decision Tree Classifier is used and applied to the signal and background data set. The cut on the classifier prediction is derived from the Figure of Merit (figure 9) which delivers the value  $c_{bdt} = 0.86$ . Dividing the number of signal events by the number of generated events returns the reconstruction efficiency  $\epsilon_s$ . The result for the muon channel is 0.20496 and therefore slightly higher than the retrieved efficiency in [1]. To obtain the branching ratio the following formula is used [1]:

$$\mathcal{B}(B \rightarrow K^* J/\psi) = \frac{N_{obs}}{\epsilon_s \cdot f_{\mu\mu} \cdot N_{B\bar{B}}}. \quad (6)$$

$N_{obs}$  is the extracted signal yield and  $N_{B\bar{B}}$  is the number of recorded B meson pairs at Belle which amounts to 770 million.  $f_{\mu\mu}$  describes the fraction of the  $J/\psi$  state that decays into muon pairs and is 5,961%.

At first, the length of the background sample is used as the value of  $N_{obs}$ . Thereby, a branching ratio of  $\mathcal{B} = 0.00194 \pm 0.000034$  is obtained. The corresponding value in [1] amounts to  $0.00124 \pm 0.00003$ .

To obtain better results different fit methods are applied. The signal component of  $M_{bc}$  is fitted with the Crystal Ball function [1]:

$$\mathcal{P}^{CB}(M_{bc}, m_0, \sigma, \alpha, n) = \begin{cases} e^{-\frac{(M_{bc}-m_0)^2}{2\sigma^2}} & \text{if } M_{bc} > m_0 - \alpha\sigma \\ \left(\frac{n}{\alpha}\right)^n e^{-\frac{\alpha^2}{2}} \left(\frac{m_0-M_{bc}}{\sigma} + \frac{n}{\alpha} - \alpha\right)^{-n} & \text{if } M_{bc} \leq m_0 - \alpha\sigma \end{cases}, \quad (7)$$

where  $m_0$  and  $\sigma$  are the mean and width of the distribution. The function describes a Gaussian that shows a power-law tail below a threshold which is defined by the parameters  $\alpha$  and  $n$  [1]. The background distribution can be approximated by the so-called ARGUS shape [1]:

$$\mathcal{P}^{ARGUS}(M_{bc}, m_0, \alpha) = M_{bc} \sqrt{1 - \left(\frac{M_{bc}}{m_0}\right)^2} e^{-\alpha(1 - (\frac{M_{bc}}{m_0})^2)}. \quad (8)$$

$\alpha$  describes the slope and  $m_0$  is the cutoff value of the distribution [1]. Moreover, an unbinned maximum likelihood fit is applied to extract the yields [1]. After that, the result can be optimized to  $0.001592 \pm 0.000016$ . Figure 11 shows the distribution of signal and background after the fit has been applied.

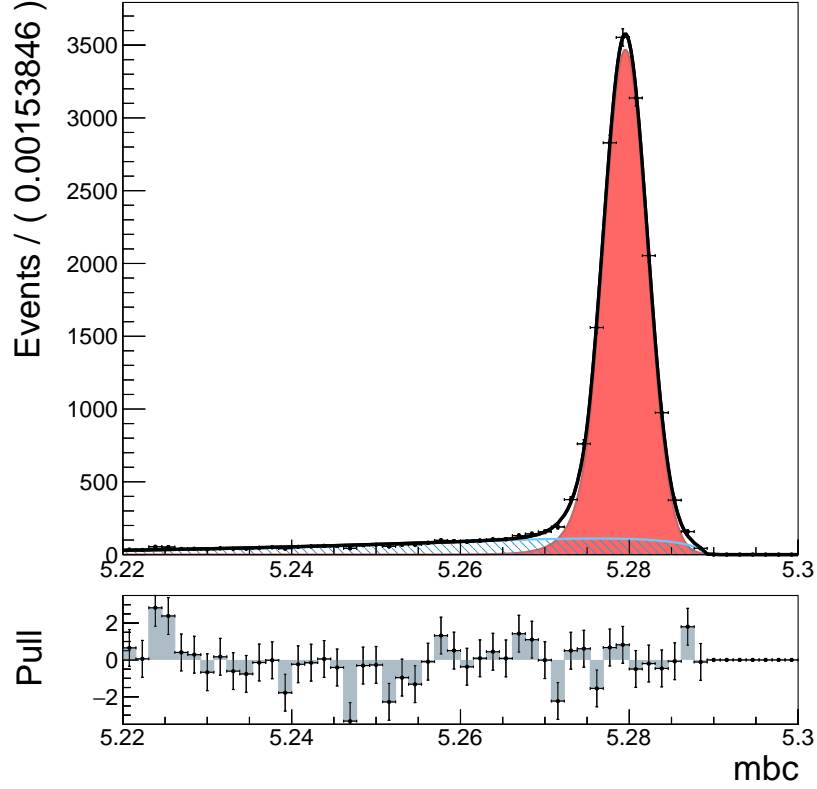


Figure 11: Distribution of signal and background for  $M_{bc}$  in the cut region  $9 \text{ GeV}^2/c^4 < q^2 < 10 \text{ GeV}^2/c^4$  after the fits 7 and 8 have been applied. The obtained result for the branching ratio is  $\mathcal{B} = 0.001592 \pm 0.000016$ .

### 3 Summary

For this Summer School project simulated Monte Carlo data has been used. Different machine learning classifiers have been tested and evaluated. Finally, the branching ratio for the decay  $B \rightarrow K^{(*)}J/\psi$  has been calculated.

## References

- [1] S. Wehle, *Angular Analysis of  $B \rightarrow K^* l l$  and Search for  $B^+ \rightarrow K^+ \tau \tau$  at the BELLE Experiment*, PhD Thesis, University of Hamburg, 2016.
- [2] J.-F. Krohn,  *$K_L^0$  Identification Studies for Belle II*, Master's Thesis, University of Hamburg.

## A Overview of Training Variables

Variable	Definition	Used in final training set?
mass	particle mass	no
$p_{tot}$	total momentum of the candidate	no
$\Delta E$	$\Delta E = E_B - E_{Beam}$	yes
Q Value	energy difference betw. mother and decay particles	no
$M_{bc}$	$M_{bc} = \sqrt{E_{Beam}^2 -  \vec{p}_B ^2}$	no
e	energy	no
Ch0_m	$K^*$ mass	yes
$\cos\Theta_B$	cosine of angle betw. B candidate and beam direction	yes
$\chi^2$	$\chi^2$ value of the vertex fit of the daughters	yes
$\Delta z_{ll}$	distance betw. the two leptons in z direction	yes
$qr_{NN}$	result of neural network flavor algorithm	yes
$qr_{LR}$	result of multidimensional likelihood ratio flavor tagger	yes
$M_{miss}$	missing mass of event	yes
$E_{vis}$	visible energy of the event	yes
mom_dir_dev	momentum direction deviation	yes
$d_{IP}$	distance to interaction point	yes
Ch0_pt	$p_{tot}$ of child 0	yes
cs-[...]	multiple variables used for continuum suppression	yes

Table 2: Overview of different variables used for the machine learning training.