Iskanje nove fizike z variacijskimi avtoenkoderji

Matic Debeljak

31. avgust 2021

1 Uvod

Po več desetletjih intenzivnih raziskav se je zgodba Standardnega Modela (SM) fizike osnovnih delcev leta 2012 z odkritjem Higgsovega bozona lepo zaokrožila. S tem so bile odkrite vse napovedane prostostne stopnje. Raziskave seveda še potekajo, saj vseh lastnosti SM še nismo postavili na preizkušnjo, kljub temu pa se SM različnim testom znova in znova zoperstavlja. Nekaj ključnih lastnosti pa mu vseeno manjka, najbolj očitne so nevtrinske mase, za katere iz meritve nevtrinskih oscilacij vemo, da so neničelne, v nasprotju z napovedjo SM. Povedano drugače, ne razumemo mehanizma, preko katerega nevtrini dobijo svojo (majhno) maso, v nasprotju z ostalimi fermioni SM, pri katerih smo vedno bolj prepričani v Higgsov mehanizem. Poleg tega je zelo očiten problem tudi temna snov, ki v SM nima kandidata. Med konceptualnimi problemi se znajdejo problem hierarhije, problem okusa, problem močne simetrije CP, in drugi.

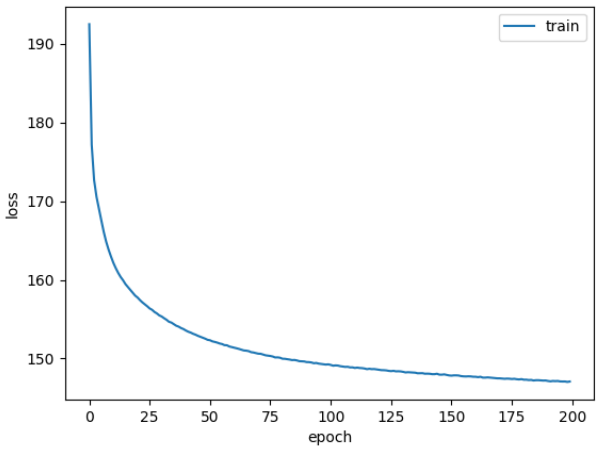
Metode strojnega učenja kažejo velik potencial za uporabo prav v iskanju nove fizike v kontekstu visoko-energijskih trkalnikov. Ker ne vemo, kakšen je signal v resnici, se naravno ponujajo metode nenadzorovanega učenja. Le-te so dandanes predmet intenzivnih raziskav. Primer, ki si ga bomo ogledali tukaj, temelji na LHC Olimpijadi. Iskali bomo dvocurkovni signal v morju dogodkov ozadja, tako da bomo ustvarili klasifikator, ki bo deloval na podlagi latentnega prostora Variacijskega AvtoEnkoderja (VAE)

2 Rezultati

Za spoznavanje delvanja VAE smo se najprel lotili analize baze ročno napisanih števk MNIST. Nato pa smo pridobljeno znanje uporabili za analizo dvocurkovnega signala.

2.1 Baza števk MNIST

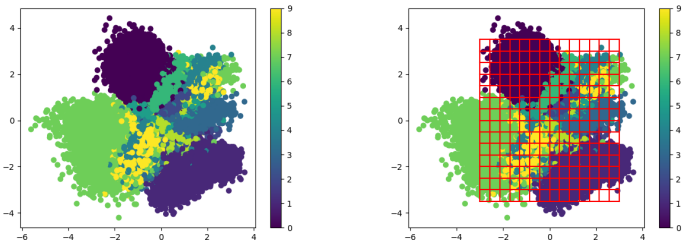
S pomočjo navodil sem implementrial enostaven model VAE in ga natreniral na bazi števk MNIST. Na sliki 1 lahko vidimo, kako se je skozi učenje spreminjala funkcija izgube.



2

Slika 1: Prikaz funkcije izgube tekom učenja na bazi števk MNIST

Kot pričakovano je funkcija izgube tekom učenja padala, VAE bi verjtno lahko učili še nekaj epoch in prišli do boljši rezultatov, saj se je funkcija izgube še kar manjšala. A model je napovedoval že dovolj dobro, zato sem se odločil da zaustavim učenje. Poglejmo si kako izgleda latentni prostor za ta primer.



Slika 2: Prikaz latentnega prostora enkoderja (levo) in mreže za katero so na sliki 3 prikazane generirane številke s pomočjo dekoderja (desno)

Poglejmo si kakšne številke generira dekoder iz latentnega prostora (slika 3)

Slika 3: Generirane številke iz latenega prostora

Na sliki 3 lahko precej jasno vidmo, kako se številke prelivajo ena v drugo. Nekaj številk je jasno ločenih od drugi (0, 1, 3, 7, 8). Medetem ko nekatere druge težje ločimo med sabo (2, 4, 5, 6, 9).

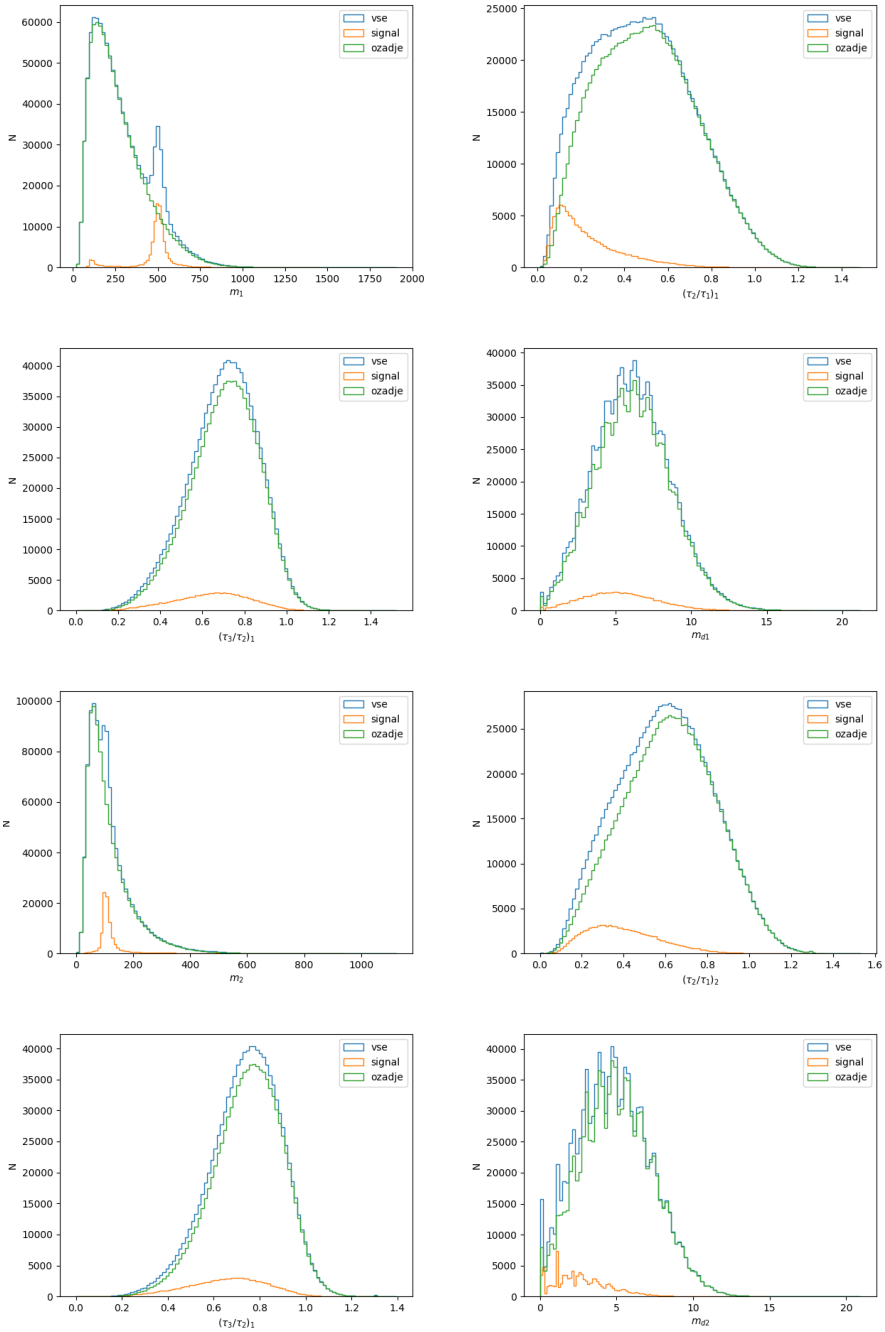
2.2 Dvocurkovni siganl - testni podatki

Po tem ko smo sestavili preprosti VAE, se lahko sedaj lotimo težjega primera. Na voljo smo imeli dve skupini podatkov: testni primer, za katerega smo imel na voljo tudi oznake in primer črne škatle, kjer oznak ni bilo podanih. Najprej smo se seveda lotili tesntega primera, ki je bolj primeren za učenje.

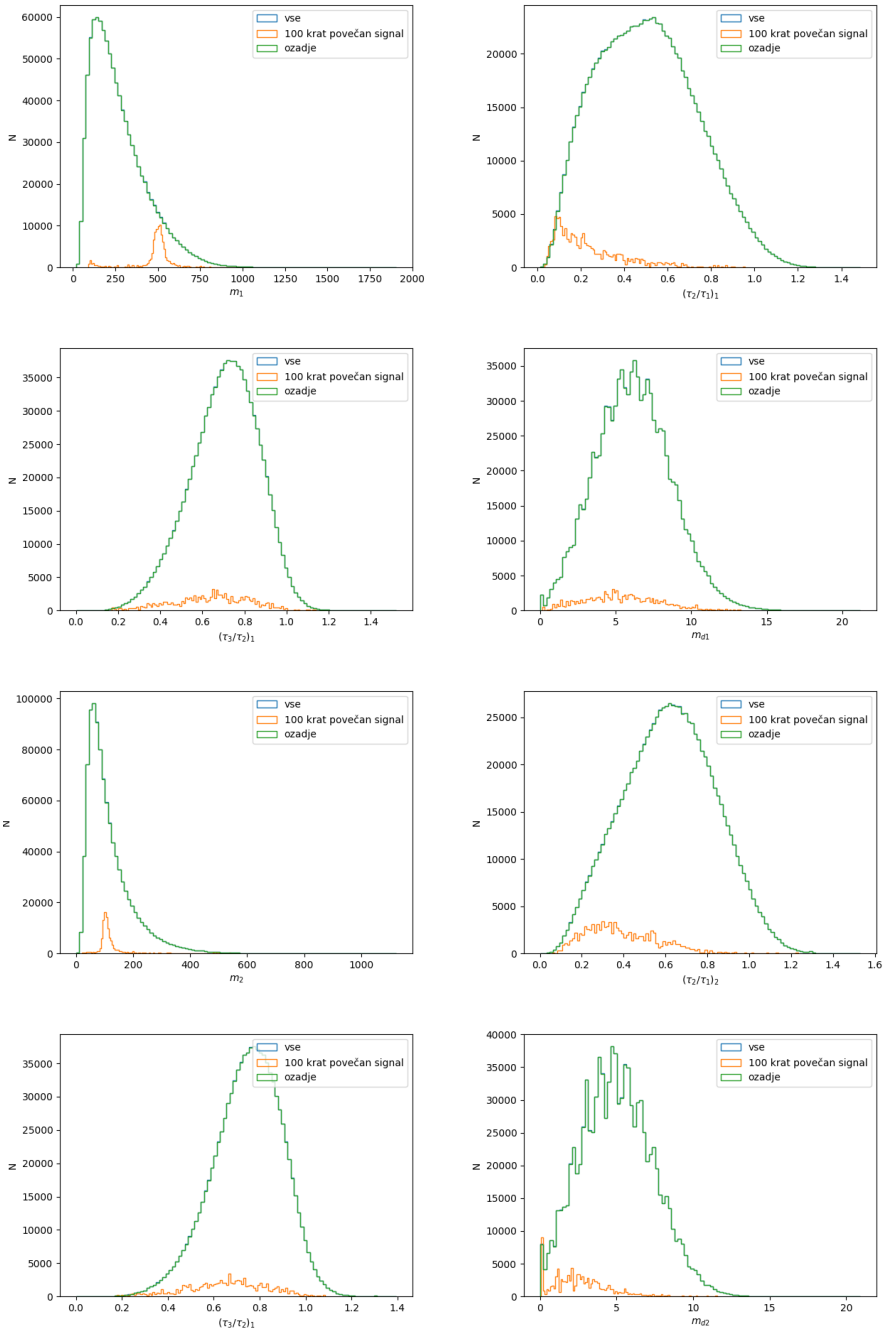
Testni podatki, so spet vsebovali dva seta podatkov. Prvi je vseboval 10% signala, drugi pa samo 0.1% signala. Vsak dogodek je bil opisan z osmimi spremeljivkami: dve invariantni masi hadronskih curkov (dveh z najvišjim p_T) m_{j1}, m_{j2} ,ter 3 spremenljivke curkovne substrukture za vsak curek. Le-te označimo z τ_2/τ_1, τ_3/τ_2 ter m_d seveda za vsakega od curkov posebej. Prvi dve spremenljivki tukaj sta tako-imenovani N-subjettiness spremenljivki, merita pa kako verjetno je, da je hadronski curek nastal iz dveh osnovnih delcev

(τ_2/τ_1) in kako verjetno je, da je nastal iz treh osnovnih delcev (τ_3/τ_2). Zadnja spremenljivka je vsota tako-imenovanih massdrop-ov, to je največjih razmerij mas hčerinskih in materinskih delcev v celotni verigi gručenja curkov. Na slikah 4 in 5 vidimo prikaz teh spremeljivk za set podatkov z 10% signalom in za set podatkov z 0.1% signalom (signal je v tem primeru ojačan, da ga sploh lahko vidimo).

Na slikah 4 in 5 vidmo, da spremelnjivke niso razporejene zelo enakomerno, kar zna predstavljati težave pri slikanju v latentni prostor in posledično pri učenju VAE, zato na podatkih uporabimo transformacijo, ki le te spremeni v bolj enakomerno porazdeljene in s tem tudi bolj primerne za učenje VAE.

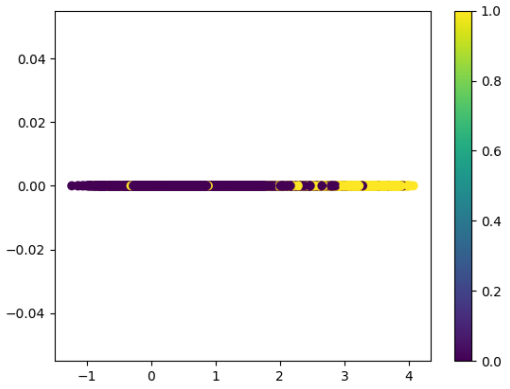


Slika 4: Prikaz spremeljivk za testni set podatkov z 10% signala



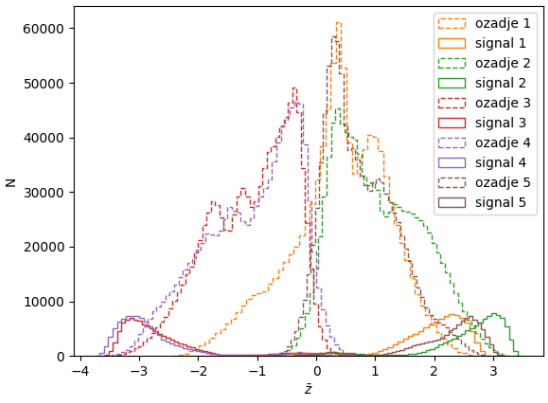
Slika 5: Prikaz spremeljivk za testni set podatkov z 0.1% signala

Ko vhodne podatke transformiram, lahko začenmo učiti model. Za oba seta podatkov sem VAE učil petkrat. Latenti prostor za set podatkov z 10% signala je prikazan na sliki 6. Vidmo, da je večina ozadja skoncentriranega na območju med -1 in 2, medtem ko signal prevaladuje na območju 3 do 4. Glede nato, da sta signal in ozadje v latentnemu prostoru precej ločena lahko predvidevam, da naš enkoder deluje dobro.

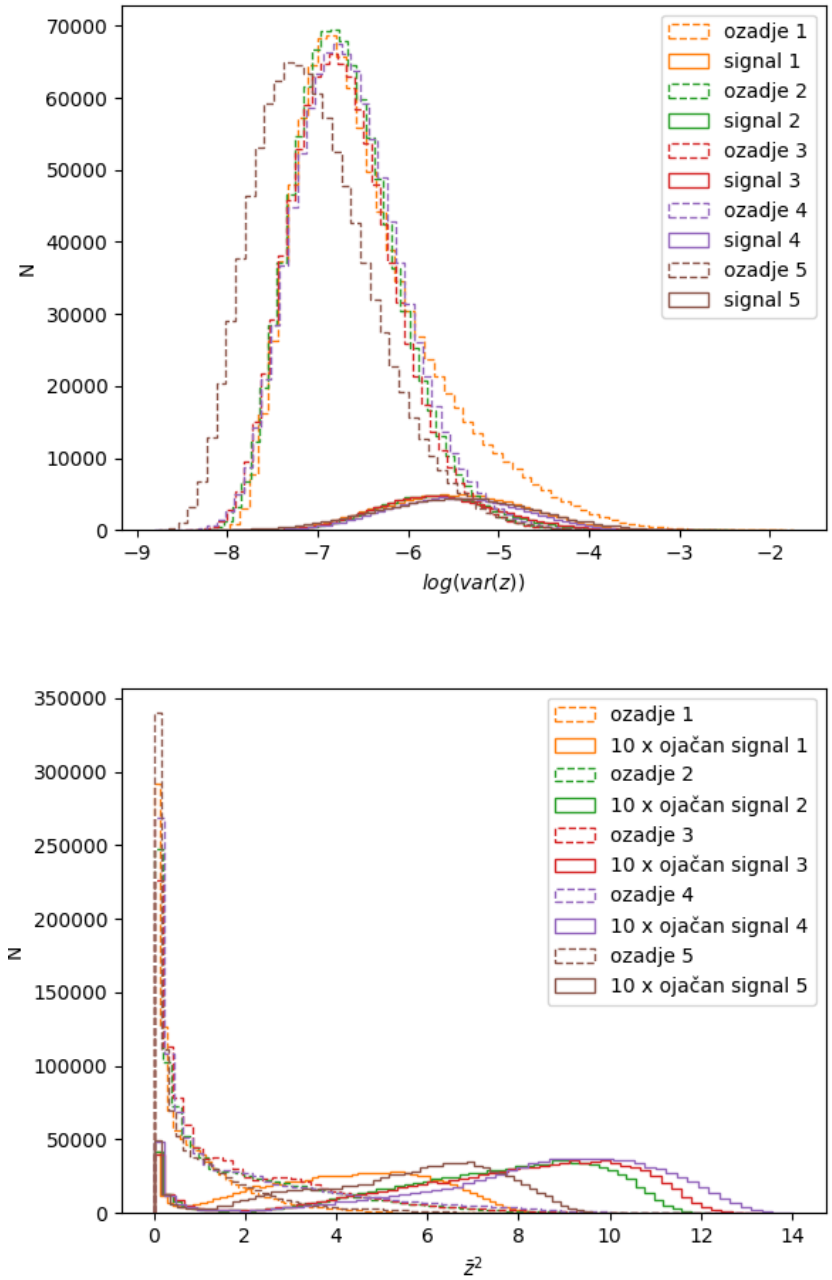


Slika 6: Latentni prostor

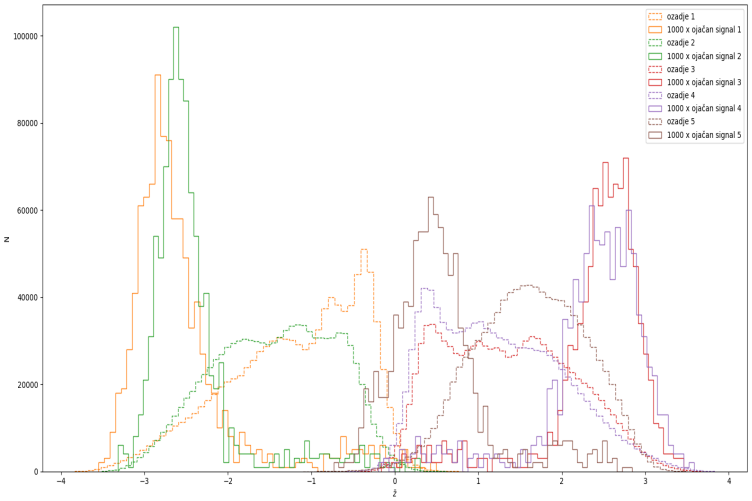
Zanima nas tudi distribucija $\bar{z}$, $\bar{z}^2$ in $\log(var(z))$ za oba seta podatkov. Prikaz na slika 7 - 10.

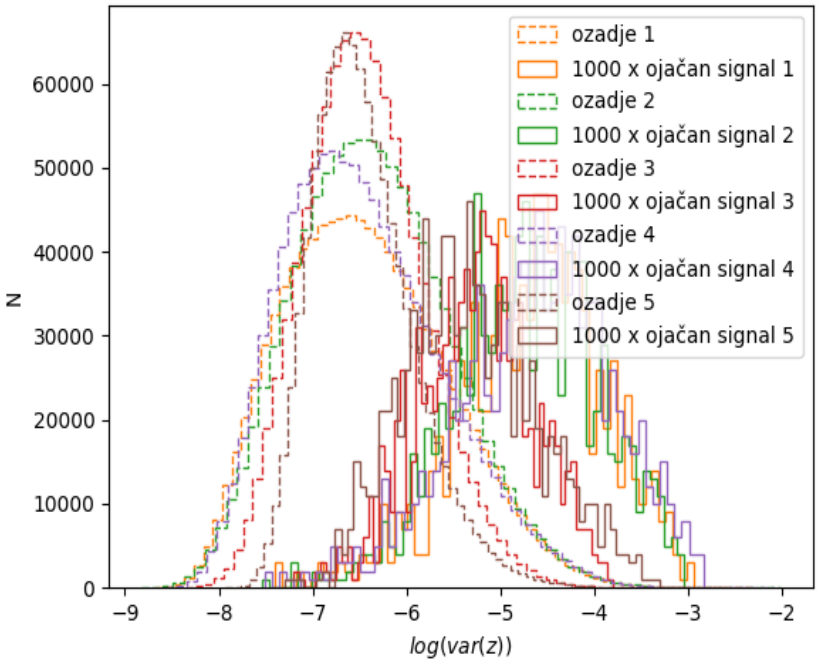


Slika 7: Distribucija $\bar{z}$ za set podatkov z 10% signalom za vseh pet ponovitev.

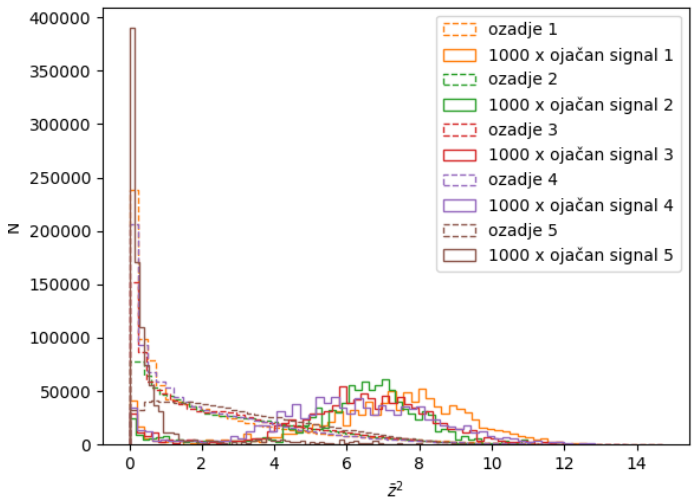


Slika 8: Distribucija $\log(var(z))$ (zgoraj) in $\bar{z}^2$ (spodaj) za set podatkov z 10% signalom za vseh pet ponovitev.





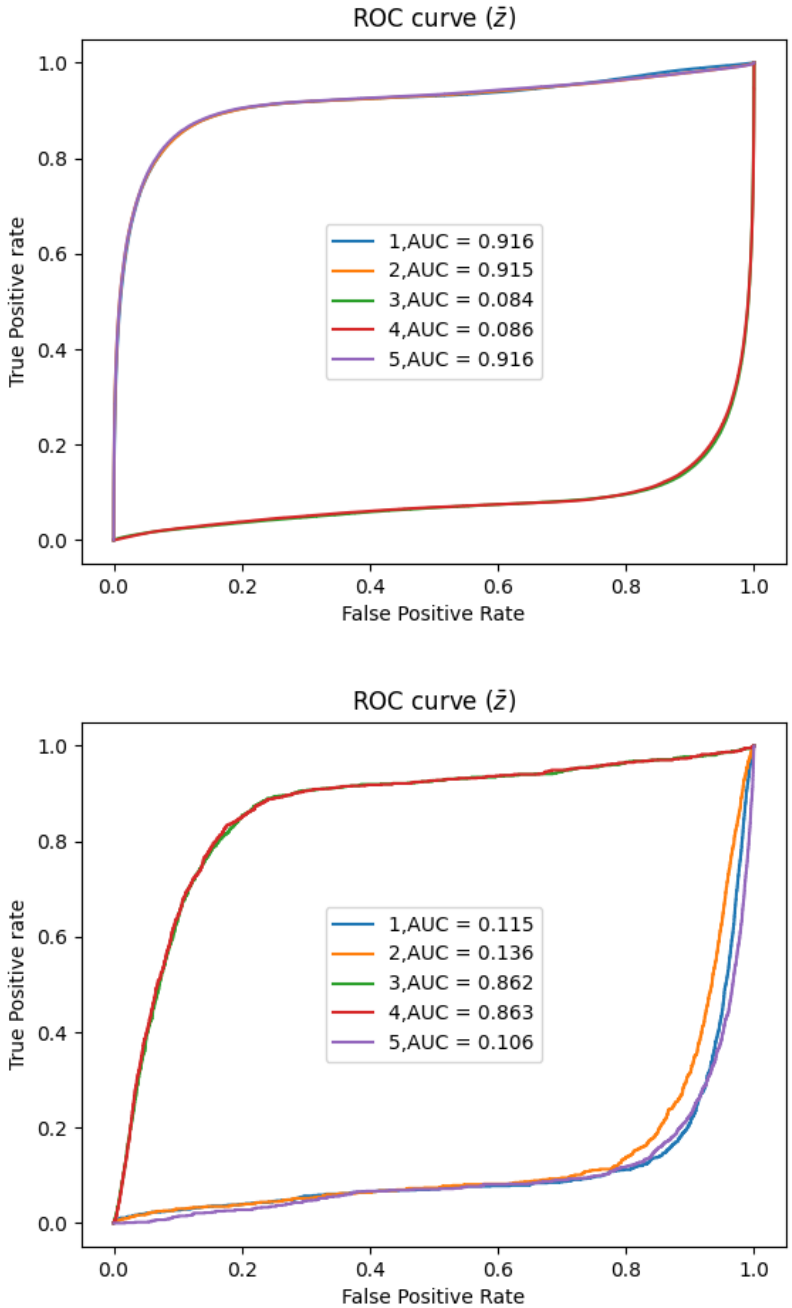
Slika 9: Distribucija $\bar{z}$ (zgoraj) in $\log(var(z))$ (spodaj za set podatkov z 0.1% signala za vseh pet ponovitev.



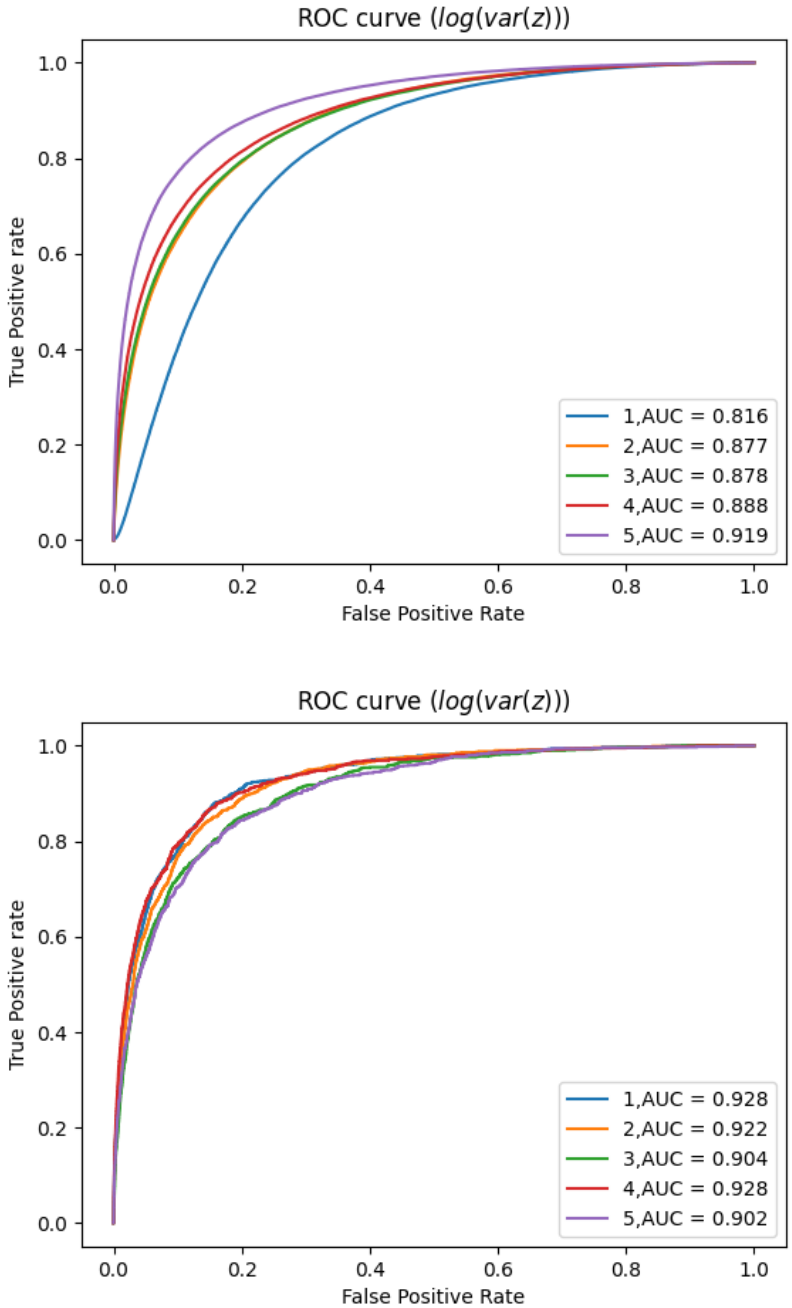
Slika 10: Distribucija $\bar{z}^2$ za set podatkov z 0.1% signala za vseh pet ponovitev.

Opazimo, da lahko signal in ozadje ločimo predvsem po $\bar{z}^2$ in $\log(var(z))$, torej lahko pričakujeo, da bota boljša kalsifikatorja kot $\bar{z}$. Na slika 11 - 14 so prikazane ROC krivulje za vse tri potencialne klasifikatorje ($\bar{z}$, $\bar{z}^2$ in $\log(var(z))$)

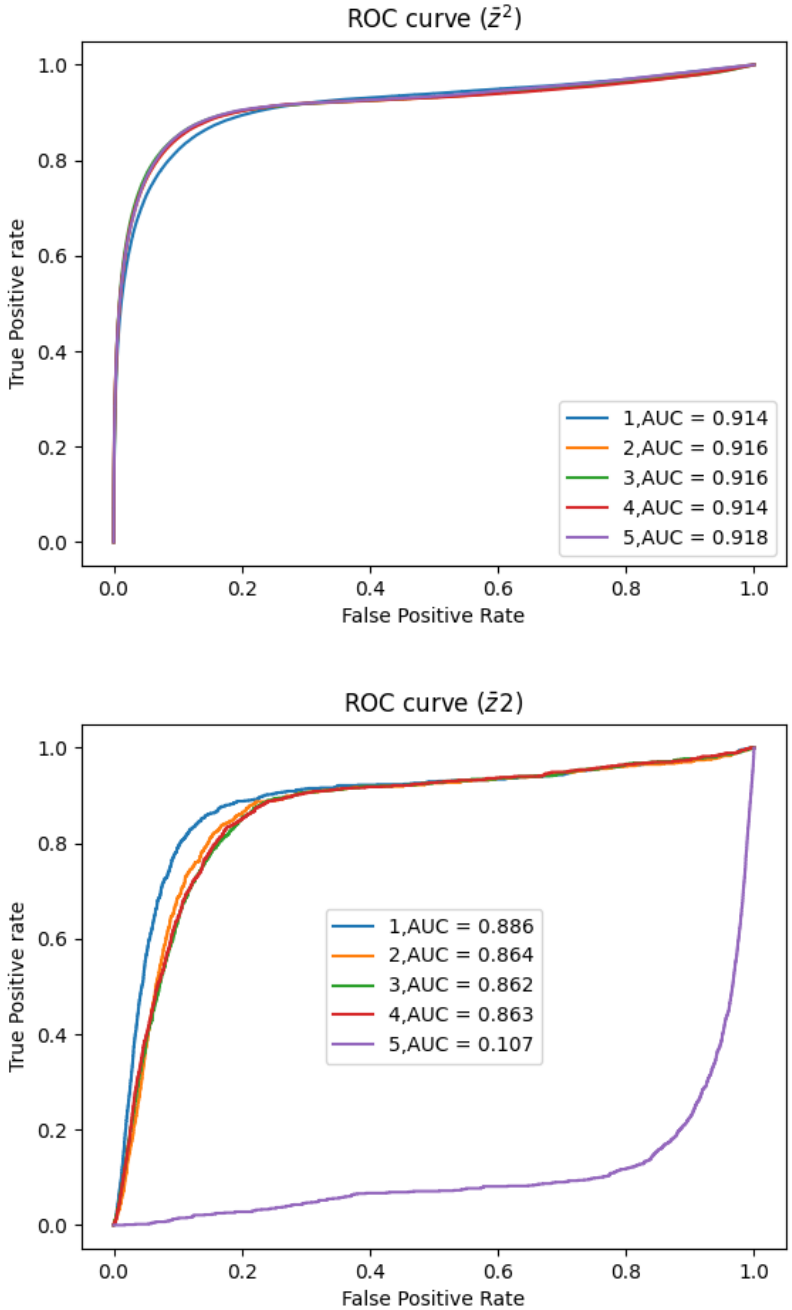
S pomočjo ROC krivulj in AUC kriterija lahko določimo najboljši kalsifikator za nadalnjo obdelavo podatkov. $\bar{z}^2$ in $\log(var(z))$ imata dokaj podobno uspešnost, jaz sem se odločil za uporabo $\bar{z}^2$.



Slika 11: ROC krivulje za podatke z 10% signala (zgoraj) in 0.1% (spodaj) ob uporabi $\bar{z}$ kot klasifikatorja

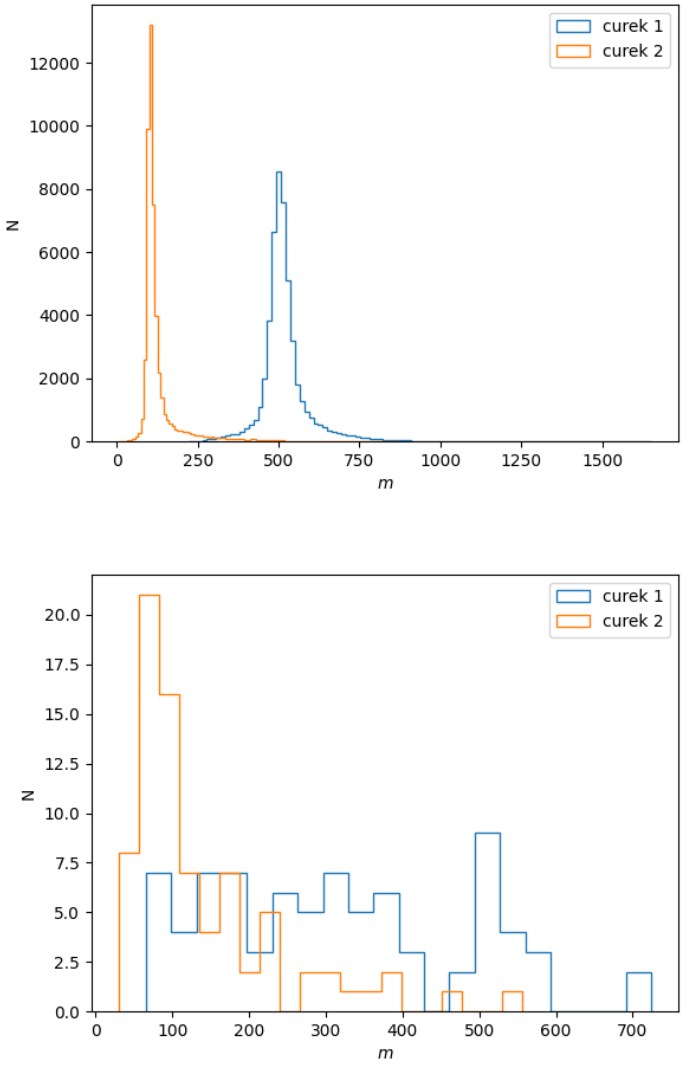


Slika 12: ROC krivulje za podatke z 10% signala (zgoraj) in 0.1% (spodaj) ob uporabi $\log(var(z))$ kot klasifikatorja



Slika 13: ROC krivulje za podatke z 10% signala (zgoraj) in 0.1% (spodaj) ob uporabi $\bar{z}^2$ kot klasifikatorja

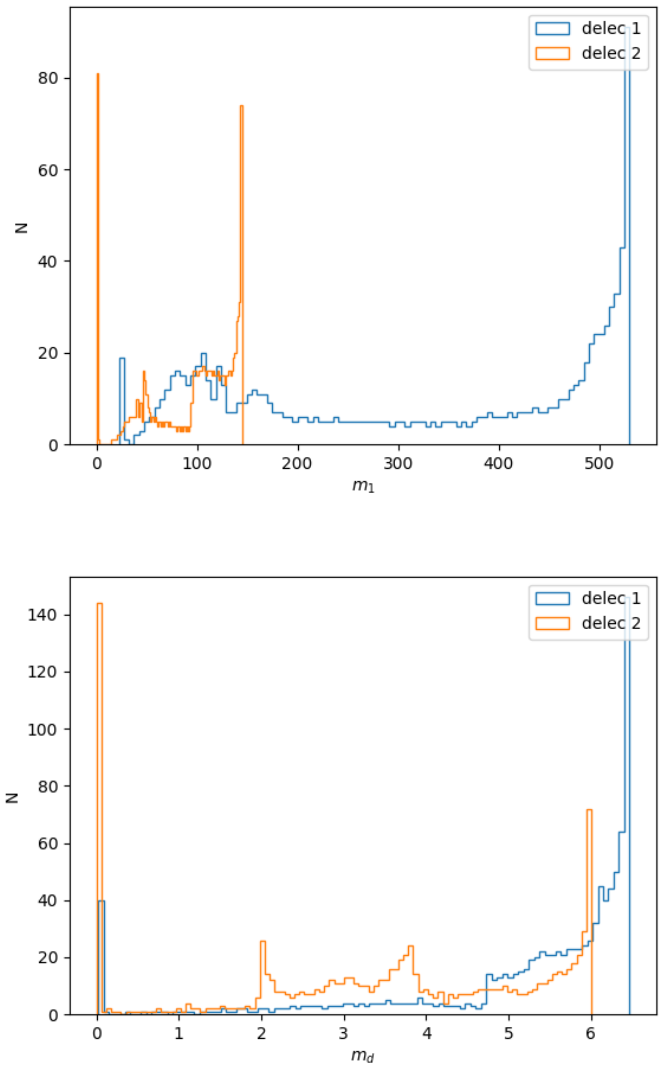
Sedaj lahko podatke sortiram po padajoči vrednosti klasifikatorja, tako bo večino signalnih dogodkov na začetku seznama. To nam pomaga pri določanju mas delecev 1 in 2 (slika 14 in 15).



Slika 14: Masi signalnih delcev za primer 10% (zgoraj) in 0.1% (spodaj)

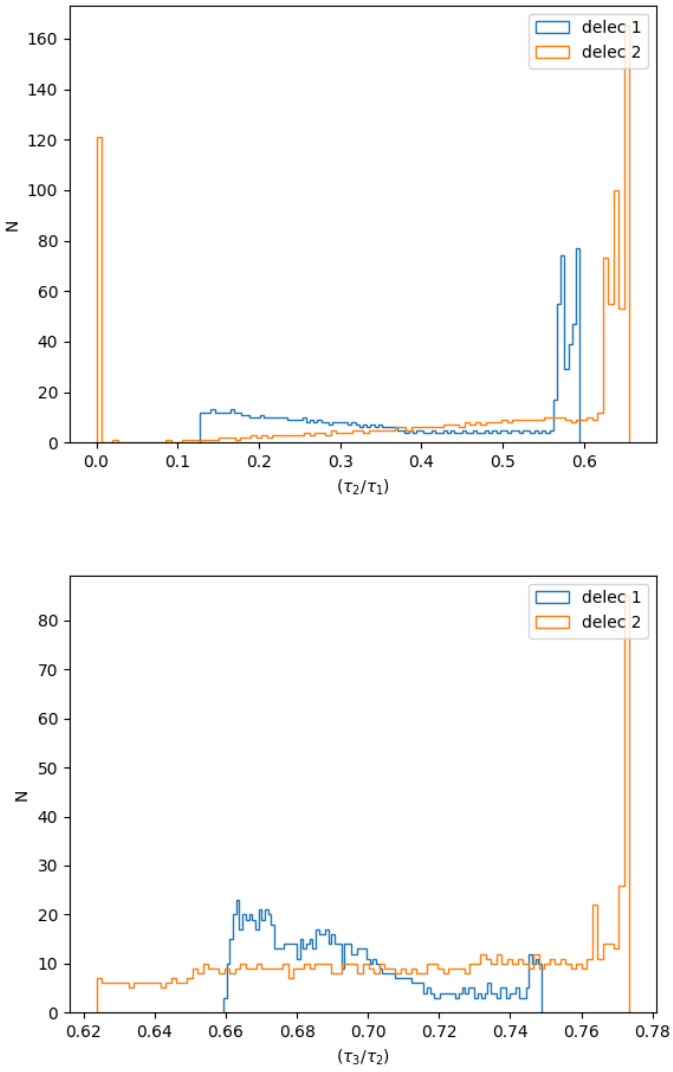
Kot pričakovano je signal precej lažje opaziti v podatkih, ki ga vsebujejo več. V primeru z 10% signala lahko določimo $m1 = 500$ GeV in $m2 = 110$ GeV. Medtem, ko v primeru z 0.1% signala te masi precej težje določimo, predvsem za curek 1.

Še ena koristna stvar, ki je počnejo VAE je generacija novih podatkov. Pogledamo si kako izgleda distribucija 8 spremeljivk generiranih podatkov (sliki 15 in 16). Generiram sem 1000 podatkov, ki so bili v latentnemu prostoru enakomerno porazdeljeni med -1 in 4.



Slika 15: Distribucija m in m_d za generirane podatke

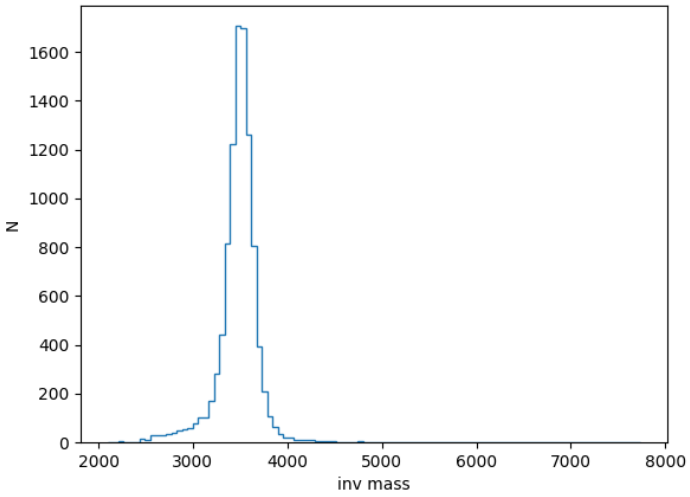
Vidimo, da grafi distribucij generiranih podatkov zgledajo drugač, kot grafi za podane podatke. To je poseldica, da sem jaz podatke generiral enakomerno po celotnem latentnem prostoru, medtem ko so bili podani podatki



Slika 16: Distribucija τ_2/τ_1 in τ_3/τ_2 za generirane podatke

zgoščeni v območju ozadja in redkejši v območju signala

S pomočjo datoteke s podatki o invarianti masi lahko določimo še invarianto maso signalnih podatkov (slika 17) kot 3.5 Tev

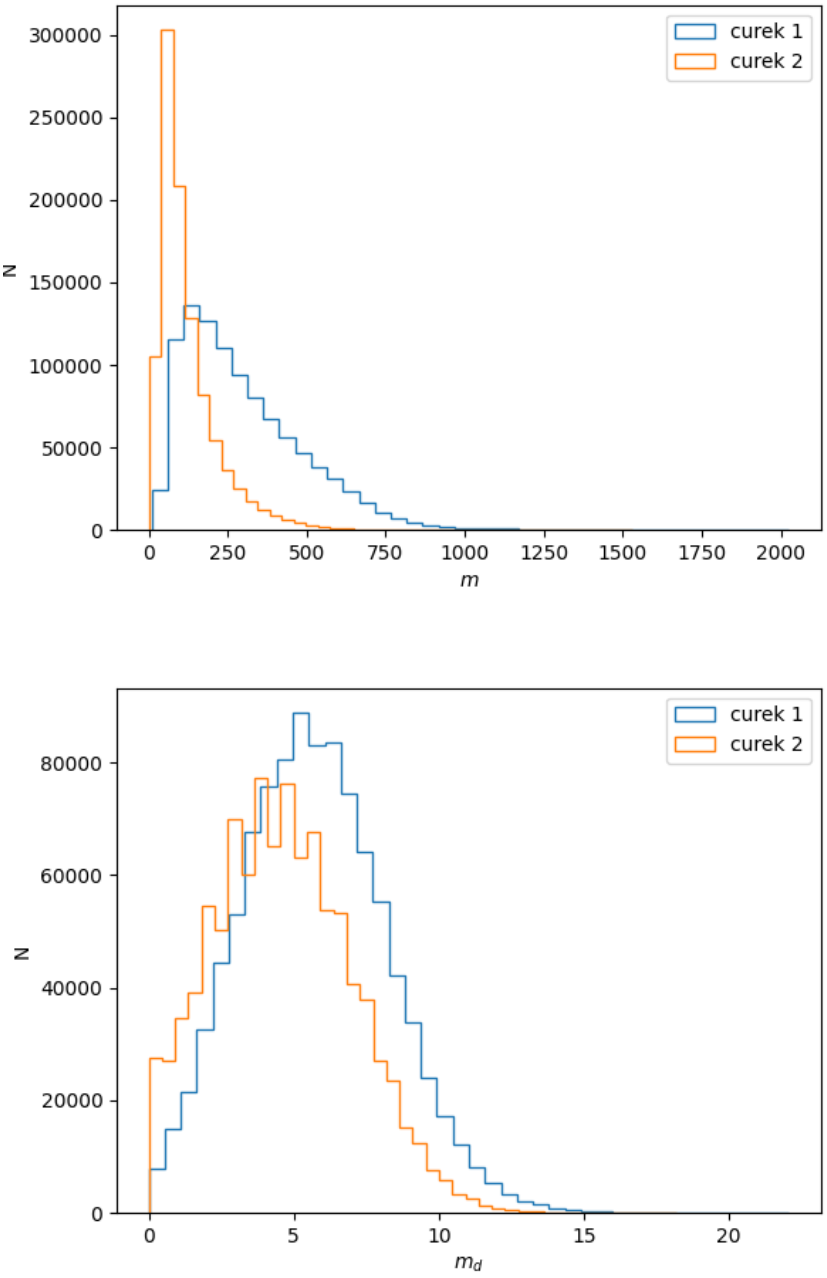


Slika 17: Distribucija invariantne mase za siganal

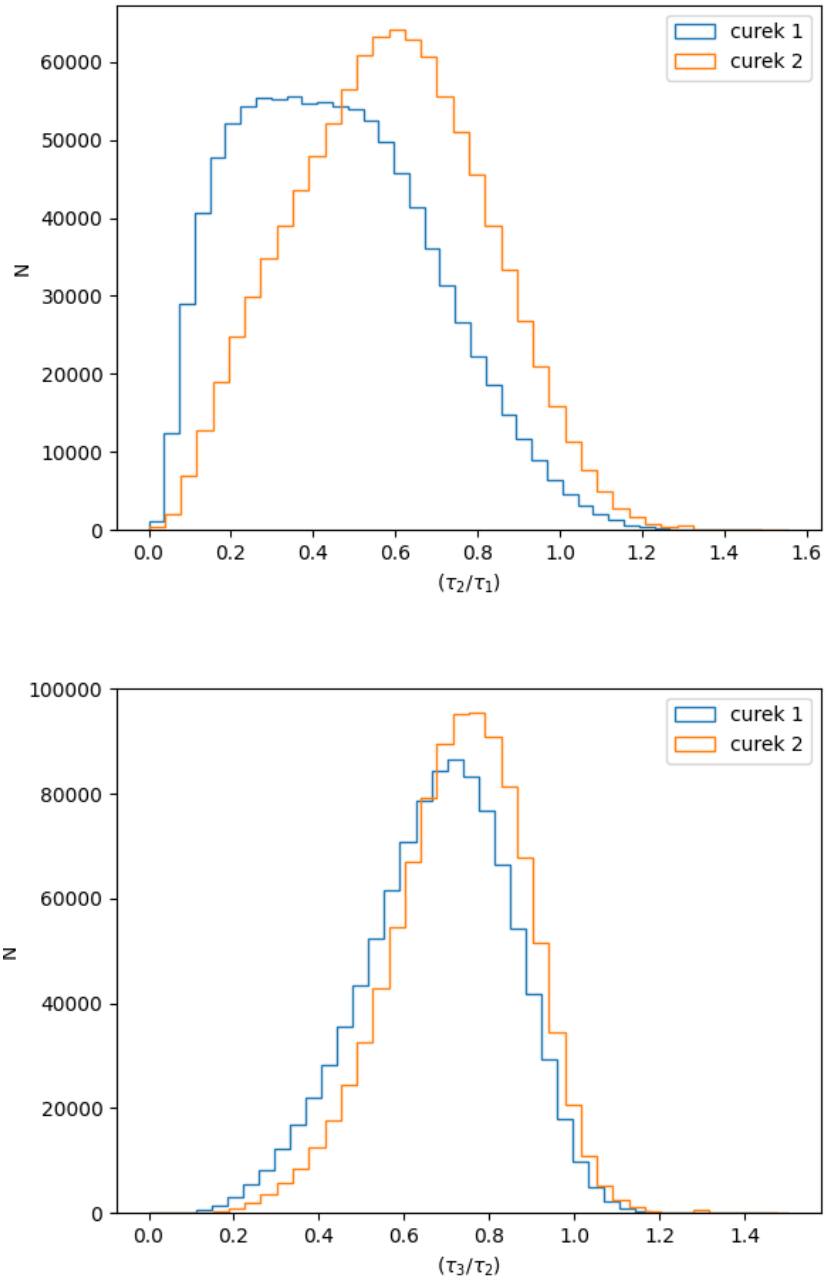
2.3 Črna škatla

Po uspešni analizi testnih podatkov se lahko lotimo še anlize "črne škatle" za katero nimamo podanih oznak za ozadje in signal. Naša naloga je torej da poskusimo izločiti signl iz ozadja.

Za začetek si ponovno pogeljmo distribucijo podatkov(sliki 18 in 19), morda bomo podobno kot v primeru s testnimi podatki z 10% signala že v distrubiciji mase opazili kakšen izrazit vrh, ki nam bo pomagal pri določitvi mase delca.

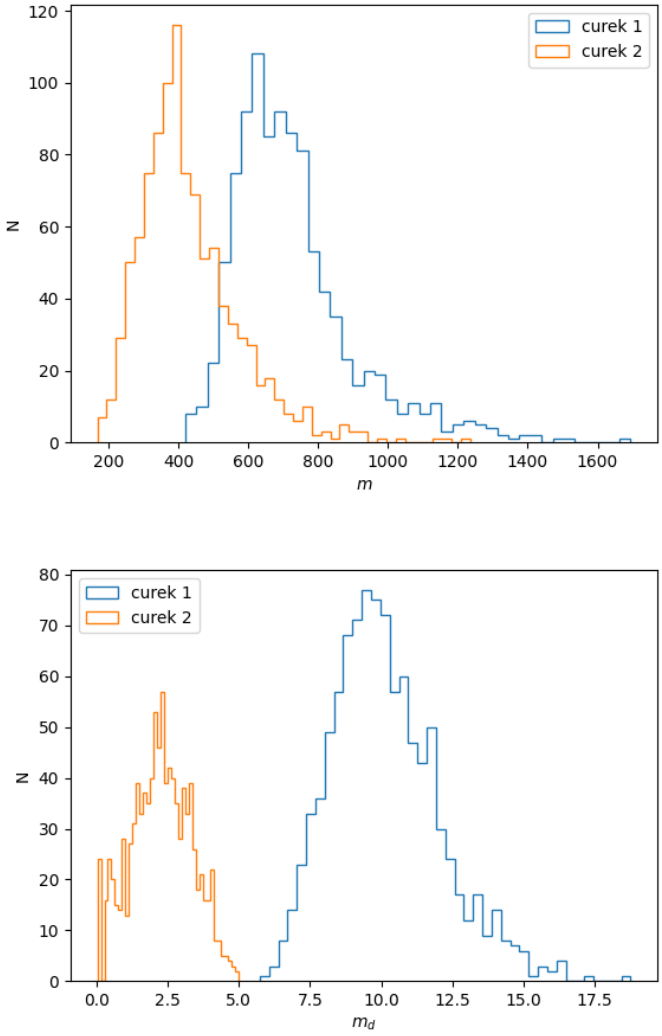


Slika 18: Distribucija m in m_d za podatke črne škatle

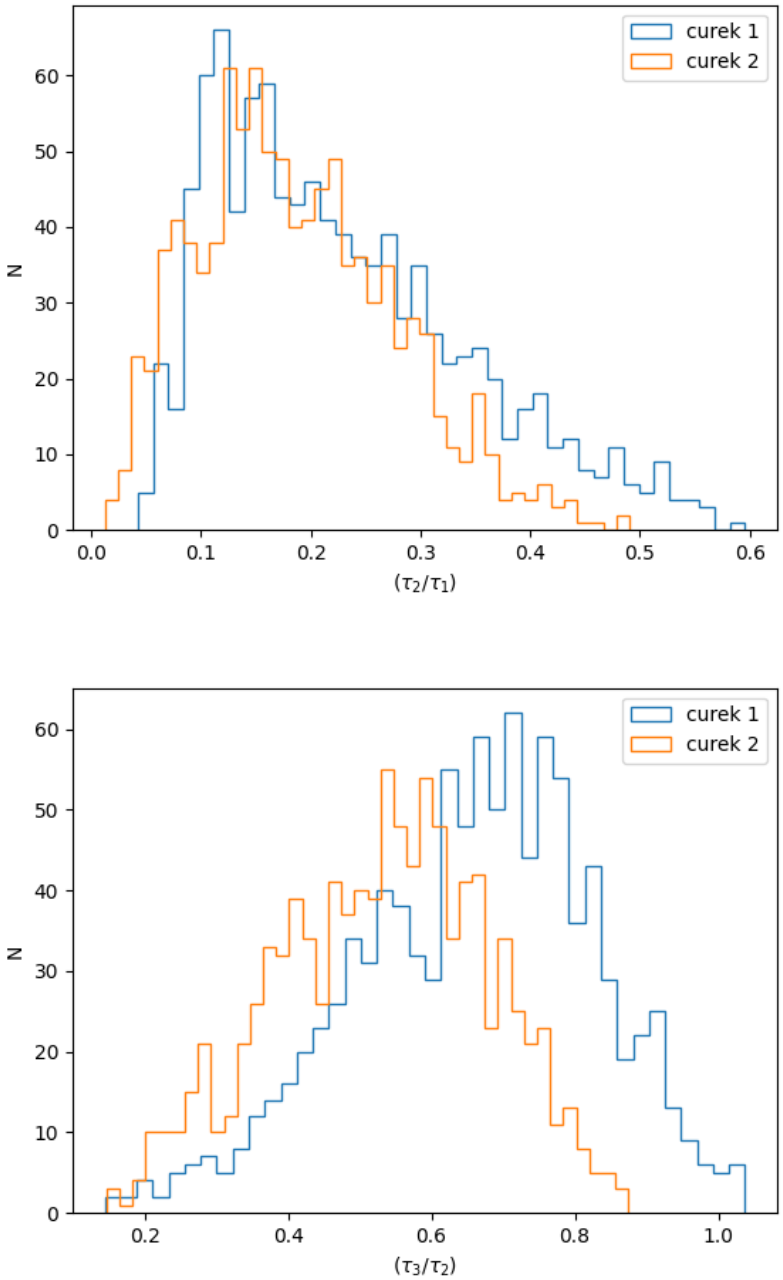


Slika 19: Distribucija τ_2/τ_1 in τ_3/τ_2 za podatke črne škatle

V distribuciji spremenljivk ne opazimo nič posebnega, zato se moramo lotiti analize podobno kot prej. Poglejmo ali lahko ločimo signal od ozadja. Na slikah 20 in 21 je prikzan filtriran signal.

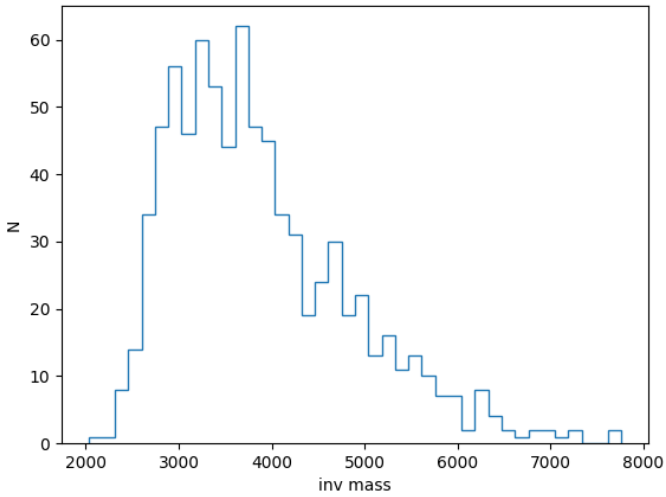


Slika 20: Distribucija m in m_d za signal iz črne škatle



Slika 21: Distribucija τ_2/τ_1 in τ_3/τ_2 za signal iz črne škatle

Podobno kot v primeru s testnimi podatki, sem podatke sortiral po padajoči vrednosti klasifikatorja in narisal prvih 1000, ter tako poskušali izolirati signal od ozadja. S pomočjo slike 19 lahko preberemo maso delca 1 in delca 2 ($m_1 = 630$ GeV, $m_2 = 400$ GeV). Nakoncu določimo še invarianto maso s pomočjo slike 22 (3.8 TeV)



Slika 22: Distribucija invariantne mase za signal iz črne škatle.

3 Zaključek

Pri tej nalogi smo se spoznali z uporabo variacijskega autoenkoderja. Uporabimo ga lahko za nenadzorovano učenje, generacijo podatkov in klasifikacijo. Naloga mi je bila všeč, saj smo najprej naredili enostaven primer, ki nam je olajšal razumevanje kompleksnejše fizikalne naloge.