

Univerzitet u Kragujevcu
Fakultet inženjerskih nauka



Seminarski rad iz predmeta: Veštačka inteligencija

Tema:

Prepoznavanje vrste biljke Irida na osnovu poznatog skupa podataka

Student:
Nikola Mitrevski 603/2017

Predmetni profesor:
dr Vesna Ranković
Predmetni asistent:
Tijana Šušteršić

Kragujevac 2021

Sadržaj:

1	Uvod	2
2	Opis korišćenja aplikacije.....	5
3	Opis delova programa sa samim izvornim kodom	8
3.1	Uvod	8
3.2	Datoteka „NaiveBayesAlg.cs“	9
4	Literatura	11

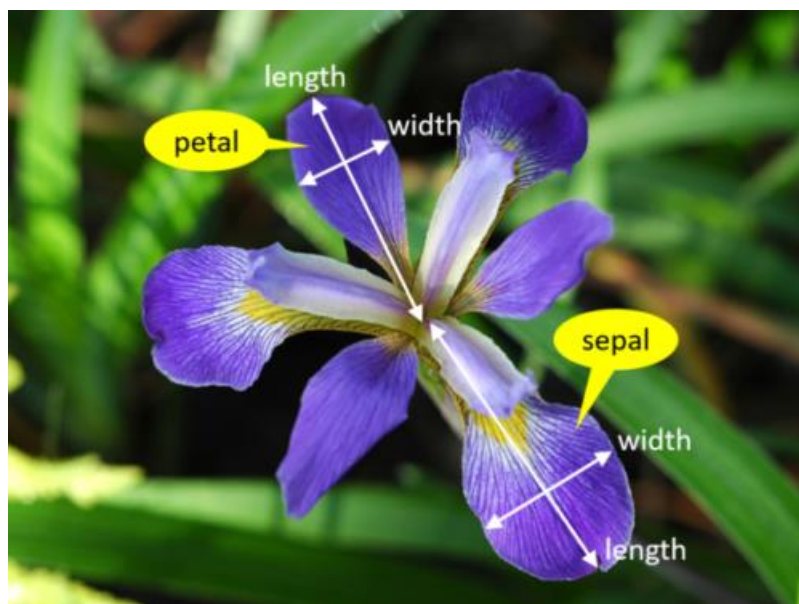
1 Uvod

Zadatak ovog projekta je pravljenje aplikacije za prepoznavanje vrste biljke *Irisa* na osnovu poznatog skupa podataka, koji je poznat kao „Fisher's Iris data set“.

Ovaj skup podataka sadrži tri klase od po 50 primeraka, pri čemu se svaka klasa odnosi na vrstu biljke *Irisa*.

Svaki primerak se sastoji od(slika 1):

- 1) dužine sepal-a(cm);
- 2) širine sepal-a(cm);
- 3) dužine petal-a(cm);
- 4) širine petal-a(cm);
- 5) klase(*Iris Setosa*, *Iris Versicolour*, *Iris Virginica*).



Slika 1 Prikaz biljke *Iris* s objašnjenjima

Ovaj projekat je baziran na „Naive Bayes“ klasifikatoru.

Naive Bayes klasifikator je zasnovan na „Bayes“ teoremi i on je namenjen za klasifikaciju podataka, odnosno svrstavanje podataka po klasama.

Bayes-ova formula za opšti slučaj glasi:

$$P(y_i | x_1, x_2, x_3, \dots, x_n) = \frac{P(y_i) \cdot P(x_1, x_2, x_3, \dots, x_n | y_i)}{\sum_{j=1}^n P(y_j) P(x_1, x_2, x_3, \dots, x_n | y_j)}$$

Slika 2 Bayes-ova formula za opšti slučaj

gde su:

- atributi X_1, X_2, \dots, X_n , klase Y_i ;
- vrednosti atributa x_1, x_2, \dots, x_n .

Naive Bayes klasifikator uvodi dve „naivne“ pretpostavke nad atributima:

- 1) svi atributi su priori podjednako važni;
- 2) svi atributi su statistički nezavisni (vrednost jednog atributa nam ne govori ništa o vrednosti drugog atributa).

Iz naivnih pretpostavki nad atributima, sledi:

$$P(x_1, x_2, x_3, \dots, x_n | y_i) = \prod_{k=1}^n P(x_k | y_i)$$

Slika 3 Formula za računanje funkcije izvesnosti

Postoje više načina za izračunavanje funkcija izvesnosti, a jedan je primena „Gausove funkcije raspodele verovatnoće“.

Gausova funkcija raspodele verovatnoće za opšti slučaj glasi:

$$P(X_i = x | y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp\left(\frac{-(x - \mu_{ik})^2}{2\sigma_{ik}^2}\right)$$
$$\mu_{ik} = \frac{1}{n} \sum_{i=1}^n x_{ik}$$
$$\sigma_{ik} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \mu_{ik})^2}$$

Slika 4 Formule za računanje Gausove funkcije raspodele verovatnoće, srednje vrednosti i standardne devijacije

gde je μ srednja vrednost, a δ standardna devijacija.

Kada se izračunaju verovatnoće klasa, jedan od načina klasifikacione odluke (svrstavanje podatka u određenu klasu, koja je za njega najverovatnija) je biranje maksimalne izračunate verovatnoće.

U ovom projektu za izračunavanje verovatnoća klasa i donošenja klasifikacione odluke će se koristiti načini koji su iznad navedeni.

Ovaj projekat se sastoji od četiri glavna dela:

- 1) razvrstavanje uzoraka po klasama;
- 2) računanje statističkih podataka za svaku kolonu, svake klase;
- 3) računanje verovatnoće za svaku klasu pomoću statističkih podataka i
- 4) maksimiziranje verovatnoća.

Uzorke je potrebno razvrstati po klasama, zbog toga što se verovatnoće računaju po klasama.

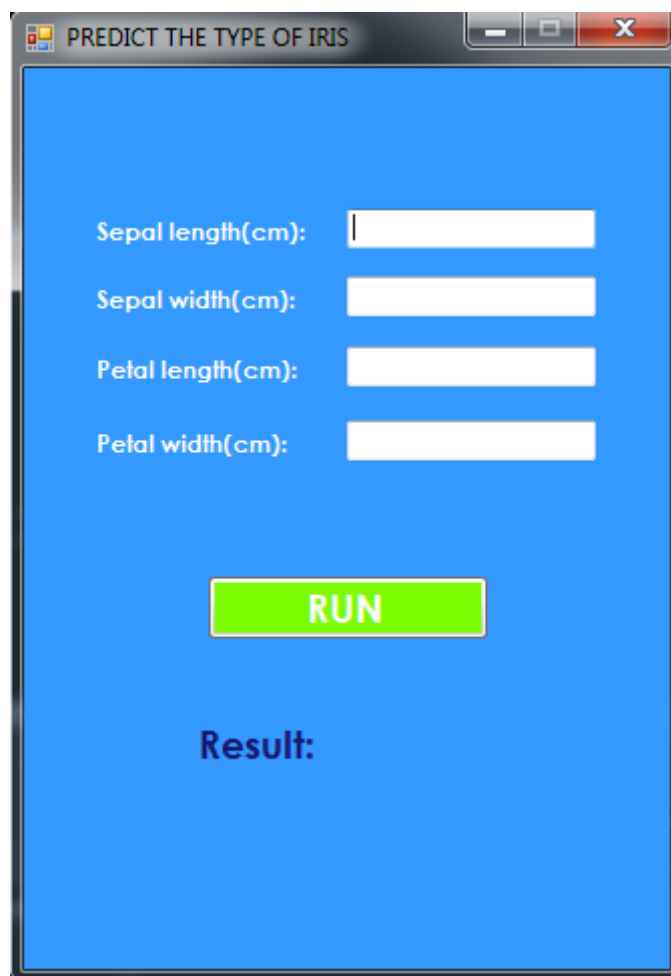
Statistički podaci koji se izračunavaju za svaku kolonu, svake klase su: srednja vrednost i standardna devijacija(odstupanje) i oni se koriste prilikom računanja verovatnoća za svaku klasu.

Verovatnoća za svaku klasu se računa primenom Gaus funkcije raspodele verovatnoće.

Maksimiziranje verovatnoća se radi, zbog toga što se na osnovu verovatnoće sa maksimalnom vrednošću predviđa o kojoj se klasi radi.

2 Opis korišćenja aplikacije


Kada korisnik pokrene aplikaciju, dobija grafički korisnički interfejs, kao na slici 5.



The image shows a graphical user interface for an application titled "PREDICT THE TYPE OF IRIS". The window has a blue background and a dark grey title bar with standard Windows window controls (minimize, maximize, close). Inside the window, there are four input fields for user data, each preceded by a label: "Sepal length(cm):", "Sepal width(cm):", "Petal length(cm):", and "Petal width(cm):". Below these fields is a prominent green button with the text "RUN" in white. At the bottom of the interface, the word "Result:" is displayed in a bold, dark blue font, indicating where the prediction output will be shown.

Slika 5 Prikaz grafičkog korisničkog interfejsa nakon pokretanja aplikacije

Nakon toga, potrebno je da popuni sva „TextBox“ polja, gde mora da vodi računa da umesto tačke stavlja zarez(slika 6).



PREDICT THE TYPE OF IRIS

Sepal length(cm): 2.4

Sepal width(cm): 1.2

Petal length(cm): 4.5

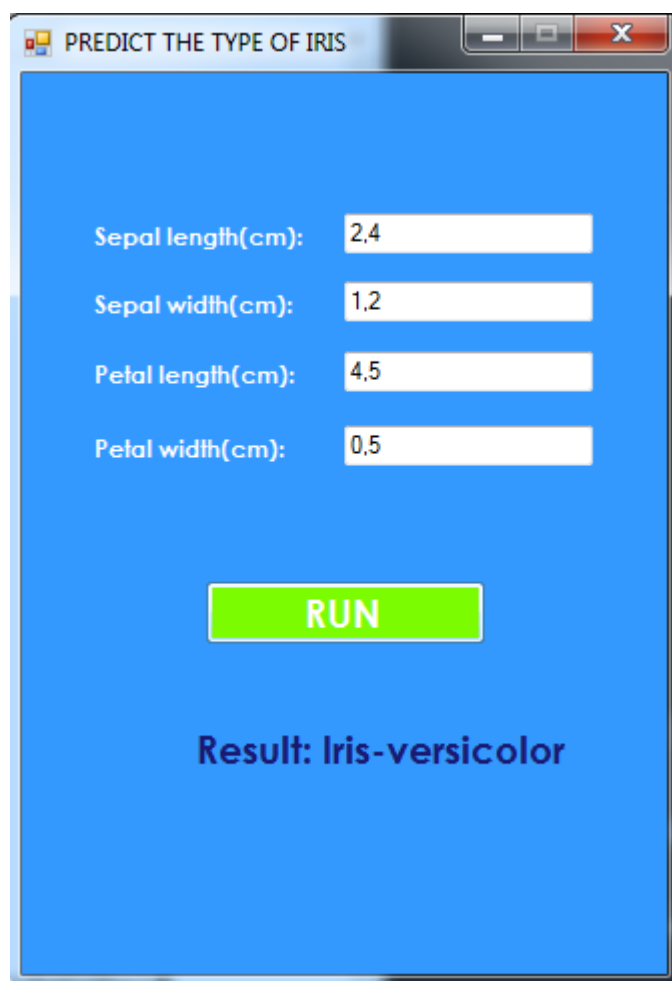
Petal width(cm): 0.5

RUN

Result:

Slika 6 Prikaz grafičkog korisničkog interfejsa nakon popunjavanja svih "TextBox" polja

Na kraju je potrebno da korisnik pritisne taster „RUN“, kako bi mu se prikazao rezultat (slika 7).



PREDICT THE TYPE OF IRIS

Sepal length(cm): 2,4

Sepal width(cm): 1,2

Petal length(cm): 4,5

Petal width(cm): 0,5

RUN

Result: Iris-versicolor

Slika 7 Prikaz grafičkog korisničkog interfejsa nakon pritiska na taster "RUN"

3 Opis delova programa sa samim izvornim kodom

3.1 Uvod

Da bi se objasnilo kako su funkcije u projektu povezane sa Naive Bayes klasifikatorom, posmatraćemo primer troklasne klasifikacije sa četiri atributa.

Krenimo od formule, sa slike 2.

Prvo što ćemo uraditi je uklanjanje delioca, zbog pojednostavljenja izračunavanja.

To je uobičajno pojednostavljenje implementacije, jer nas često više zanima predviđanje, nego verovatnoća.

U ovom primeru, formula sa slike 2, postaje:

$\begin{aligned}P(y1 x1, x2, x3, x4) &= P(y1) * P(x1, x2, x3, x4 y1) \\P(y2 x1, x2, x3, x4) &= P(y2) * P(x1, x2, x3, x4 y2) \\P(y3 x1, x2, x3, x4) &= P(y3) * P(x1, x2, x3, x4 y3)\end{aligned}$
--

Slika 8 Uprošćene Bayes formule za određivanje verovatnoća klase

U ovom projektu se formule sa slika 8, izračunavaju unutar funkcije „calculateClassProbabilities“.

Nakon toga primenimo formulu sa slike 3.

U ovom primeru, formula sa slike 3, postaje:

$\begin{aligned}P(x1, x2, x3, x4 y1) &= P(x1 y1) * P(x2 y1) * P(x3 y1) * P(x4 y1) \\P(x1, x2, x3, x4 y2) &= P(x1 y2) * P(x2 y2) * P(x3 y2) * P(x4 y2) \\P(x1, x2, x3, x4 y3) &= P(x1 y3) * P(x2 y3) * P(x3 y3) * P(x4 y3)\end{aligned}$

Slika 9 Formule za računanje funkcija izvesnosti

U ovom projektu se formule sa slika 9, izračunavaju unutar funkcije „calculateClassProbabilities“.

Zatim verovatnoće($P(x_i|y_j)$) ćemo u ovom primeru računati, primenom formule sa slike 4.

U ovom projektu formule sa slike 4 su predstavljene funkcijama: „calculateProbability“, „mean“ i „stdev“.

Nakon računanja verovatnoća, radi se maksimiziranje verovatnoća, kako bi se pronašla klasa kojoj dati uzorak pripada.

U ovom projektu maksimiziranje verovatnoća se radi pomoću funkcije „predict“.

3.2 Datoteka „NaiveBayesAlg.cs“

Ova datoteka sadrži sledeće funkcije:

- 1) Funkcija „result“ (slika 10) je prva funkcija koja se poziva iz ove datoteke i ona je namenjena za vraćanje rezultata predviđene vrste biljke Irida. Njen povratni tip je string i ona unutar sebe poziva funkcije: „loadIrisDataset“, „separateByClass“, „summarizeByClass“ i „predict“. Parametri ove funkcije su: „sepalLength“, „sepalWidth“, „petalLength“, „petalWidth“ i ovi parametri su zadati od strane korisnika s ciljem da se odredi o kojoj vrsti biljke Irida se radi.

```
public string result(double sepalLength, double sepalWidth, double petalLength, double petalWidth)...
```

Slika 10 Prikaz dela funkcije "result", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 2) Funkcija „loadIrisDataset“ (slika 11) je funkcija namenjena za učitavanje podataka iz fajla. Njen povratni tip je niz string-ova.

```
private string[] loadIrisDataset()...
```

Slika 11 Prikaz dela funkcije "loadIrisDataset", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 3) Funkcija „separateByClass“ (slika 12) je funkcija namenjena za razvrstavanje učitanih uzoraka iz fajla po klasama. Njen povratni tip je matrica string-ova. Parametar ove funkcije je „irisDataset“ i on predstavlja učitane uzorke iz fajla.

```
private string[,] separateByClass(string[] irisDataset)...
```

Slika 12 Prikaz dela funkcije "separateByClass", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 4) Funkcija „summarizeByClass“ (slika 13) je funkcija namenjena za pozivanje funkcije „summarizeDataset“. Njen povratni tip je matrica double-ova. Parametri ove funkcije su: „separated“ i „x“. Parametar „separated“ predstavlja uzorke iz fajla razvrstane po klasama, a parametar „x“ predstavlja redni broj klase.

```
private double[,] summarizeByClass(string[,] separated, int x)...
```

Slika 13 Prikaz dela funkcije "summarizeByClass", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 5) Funkcija „summarizeDataset“ (slika 14) je funkcija namenjena za pronalaženje statističkih podataka, koji su potrebni za određivanje verovatnoće za datu klasu. Statistički podaci se unutar ove funkcije određuju pozivanjem funkcija „mean“ i „stdev“ i to za svaku kolonu date klase. Njen povratni tip je matrica double-ova. Parametri ove funkcije su: „separatedDataByClass“ i „x“. Parametar „separatedDataByClass“ predstavlja uzorke iz fajla razvrstane po klasama, a parametar „x“ predstavlja redni broj klase.

```
private double[,] summarizeDataset(string[,] separatedDataByClass, int x)...
```

Slika 14 Prikaz dela funkcije "summarizeDataset", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 6) Funkcija „mean“ (slika 15) je funkcija namenjena za pronalaženje srednje vrednosti zadatih brojeva. Njen povratni tip je double. Parametar ove funkcije je „numbers“, koji predstavlja brojeve za koje se računa srednja vrednost.

```
private double mean(double[] numbers)...
```

Slika 15 Prikaz dela funkcije "mean", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 7) Funkcija „stdev“ (slika 16) je funkcija namenjena za računanje standardne devijacije zadatih brojeva. Njen povratni tip je double. Parametar ove funkcije je „numbers“, koji predstavlja brojeve za koje se računa standardna devijacija.

```
private double stdev(double[] numbers)...
```

Slika 16 Prikaz dela funkcije "stdev", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 8) Funkcija „predict“ (slika 17) je funkcija namenjena za predviđanje vrste biljke Irisa na osnovu izračunate verovatnoće. Njen povratni tip je int i ona unutar sebe poziva funkciju „calculateClassProbabilities“. Parametri ove funkcije su: „summaries1“, „summaries2“, „summaries3“, „sepalLength“, „sepalWidth“, „petalLength“ i „petalWidth“. Parametar „summaries1“ predstavlja statističke podatke za klasu „Iris-Setosa“. Parametar „summaries2“ predstavlja statističke podatke za klasu „Iris-versicolor“. Parametar „summaries3“ predstavlja statističke podatke za klasu „Iris-virginica“. Parametri: „sepalLength“, „sepalWidth“, „petalLength“, „petalWidth“ su zadati od strane korisnika s ciljem da se odredi o kojoj vrsti biljke Irisa se radi.

```
private int predict(double[,] summaries1, double[,] summaries2, double[,] summaries3,  
double sepalLength, double sepalWidth, double petalLength, double petalWidth)...
```

Slika 17 Prikaz dela funkcije "predict", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 9) Funkcija „calculateClassProbabilities“ (slika 18) je funkcija namenjena za računanje verovatnoće za svaku klasu. Njen povratni tip je niz double-ova i ona unutar sebe poziva funkciju „calculateProbability“. Parametri ove funkcije su: „summaries1“, „summaries2“, „summaries3“, „sepalLength“, „sepalWidth“, „petalLength“ i „petalWidth“. Parametar „summaries1“ predstavlja statističke podatke za klasu „Iris-Setosa“. Parametar „summaries2“ predstavlja statističke podatke za klasu „Iris-versicolor“. Parametar „summaries3“ predstavlja statističke podatke za klasu „Iris-virginica“. Parametri: „sepalLength“, „sepalWidth“, „petalLength“, „petalWidth“ su zadati od strane korisnika s ciljem da se odredi o kojoj vrsti biljke Irisa se radi.

```
private double[] calculateClassProbabilities(double[,] summaries1, double[,] summaries2, double[,] summaries3,  
double sepalLength, double sepalWidth, double petalLength, double petalWidth)...
```

Slika 18 Prikaz dela funkcije "calculateClassProbabilities", koja se nalazi u datoteci "NaiveBayesAlg.cs"

- 10) Funkcija „calculateProbability“ (slika 19) je funkcija namenjena za računanje verovatnoće, primenom „Gausove funkcije raspodele verovatnoće“. Njen povratni tip je double. Parametri ove funkcije su „x“, „mean“, „stdev“, gde je parametar „mean“ srednja vrednost za x, a parametar „stdev“ standardna devijacija za x.

```
private double calculateProbability(double x, double mean, double stdev)...
```

Slika 19 Prikaz dela funkcije "calculateProbability", koja se nalazi u datoteci "NaiveBayesAlg.cs"

4 Literatura

[1] Machine Learning UCI – Iris Data Set, link: <https://archive.ics.uci.edu/ml/datasets/Iris> (16.05.2021, 22:17)

[2] Wikipedia – Naive Bayes classifier, link: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (16.05.2021, 22:22)

[3] Moodle portal - Računarska tehnika i softversko inženjerstvo, četvrta godina, drugi semestar, Veštačka inteligencija, Naivni Bajes, link: <http://moodle.fink.rs> (16.05.2021, 22:23)