Upotreba i usporedba metoda nenadzirane i nadzirane klasifikacije SENTINEL-2 satelitskih snimki na području Grada Zagreba u programskom jeziku R

Rukavina, Dominik

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Mining, Geology and Petroleum Engineering / Sveučilište u Zagrebu, Rudarsko-geološko-naftni fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:169:086466

SVEUČILIŠTE U ZAGREBU

RUDARSKO-GEOLOŠKO-NAFTNI FAKULTET

Diplomski studij Hidrogeologija i inženjerska geologija

UPOTREBA I USPOREDBA METODA NENADZIRANE I NADZIRANE KLASIFIKACIJE SENTINEL-2 SATELITSKIH SNIMKI NA PODRUČJU GRADA ZAGREBA U PROGRAMSKOM JEZIKU R

Diplomski rad

Dominik Rukavina

GI2118

Zagreb, 2023.

KLASA: 602-01/23-01/23
URBROJ: 251-70-03-232
U Zagrebu, 16.02.2023.

Dominik Rukavina, student

RJEŠENJE O ODOBRENJU TEME

Na temelju vašeg zahtjeva primljenog pod KLASOM 602-01/23-01/23, URBROJ: 251-70-03-231 od 13.02.2023. priopćujemo vam temu diplomskog rada koja glasi:

UPOTREBA I USPOREDBA METODA NENADZIRANE I NADZIRANE KLASIFIKACIJE SENTINEL-2 SATELITSKIH SNIMKI NA PODRUČJU GRADA ZAGREBA U PROGRAMSKOM JEZIKU R

Za mentora ovog diplomskog rada imenuje se u smislu Pravilnika o izradi i obrani diplomskog rada doc.dr.sc. Ivan Medved nastavnik Rudarsko-geološko-naftnog-fakulteta Sveučilišta u Zagrebu.

Mentor:

(potpis)

doc.dr.sc. Ivan Medved

(titula, ime i prezime)

Predsjednik povjerenstva za završne i diplomske ispite:

(potpis)

Doc.dr.sc. Zoran Kovač

(titula, ime i prezime)

Prodekan za nastavu i studente:

(potpis)

Izv.prof.dr.sc. Borivoje Pašić

(titula, ime i prezime)

Sveučilište u Zagrebu
Rudarsko-geološko-naftni fakultet

Diplomski rad

UPOTREBA I USPOREDBA METODA NENADZIRANE I NADZIRANE KLASIFIKACIJE SENTINEL-2 SATELITSKIH SNIMAKI NA PODRUČJU GRADA ZAGREBA U PROGRAMSKOM JEZIKU R

Dominik Rukavina

Rad izrađen: Sveučilište u Zagrebu
Rudarsko-geološko-naftni fakultet
Zavod za geofizička istraživanja i rudarska mjerenja
Pierottijeva 6, 10 000 Zagreb

Sažetak

Klasifikacija SENTINEL-2 satelitske snimke na području Grada Zagreba pomoću metoda nenadziranog i nadziranog strojnog učenja. Prikupljanje ulaznih podataka rađeno je u softveru QGIS. Izgradnja modela nenadzirane i nadzirane klasifikacije provedena je u R-studio-u. Usporedba točnosti predikcije korištenih modela random forest i extreme gradient boost. Vizualni pregled dobivenih rezultata stvorenih u programskom jeziku R.

Ključne riječi: Klasifikacija, NDVI, SENTINEL, R, RStudio, QGIS, Zagreb, nenadzirano, nadzirano

Završni rad sadrži: 52 stranice, 4 tablice, 32 slike, 0 priloga i 31 referenca.

Jezik izvornika: Hrvatski

Pohrana rada: Knjižnica Rudarsko-geološko-naftnog fakulteta, Pierottijeva 6, Zagreb

Mentori: Dr. sc. Ivan Medved, docent RGNF

Ocjenjivači: Dr. sc. Ivan Medved, docent RGNF
Dr. sc. Jasna Orešković, izvanredni profesor RGNF
Dr. sc. Tomislav Korman, izvanredni profesor RGNF

Datum obrane: 24.2.2023., Rudarsko-geološko-naftni fakultet Sveučilišta u Zagrebu

University of Zagreb
Faculty of Mining, Geology and Petroleum Engineering

Master's Thesis

USE AND COMPARISON OF UNSUPERVISED AND SUPERVISED CLASSIFICATION METHODS OF SENTINEL-2 SATELLITE IMAGES OF ZAGREB CITY AREA IN R PROGRAMMING LANGUAGE

Dominik Rukavina

Thesis completed at: University of Zagreb
Faculty of mining, Geology and Petroleum Engineering
Institute of Geophysical Exploration and Mine Surveying
Pierottijeva 6, 10 000 Zagreb

Abstract

Classification of SENTINEL-2 satellite images of Zagreb City area using the methods of unsupervised and supervised machine learning. Input data collection was done in QGIS software. Construction of unsupervised and supervised classification models was performed in R-studio. Comparison of prediction accuracy of used random forest and extreme gradient boost models. Visual overview of the results created in R programming language.

Keywords: Classification, NDVI, SENTINEL, R, RStudio, QGIS, Zagreb, unsupervised, supervised

Thesis contains: 52 pages, 4 tables, 32 figures, 0 appendixes and 31 references.

Original in: Croatian

Archived in: Library of Faculty of Mining, Geology and Petroleum Engineering, Pierottijeva 6, Zagreb

Supervisors: PhD Ivan Medved, Assistant Professor RGNF

Reviewers: PhD Ivan Medved, Assistant Professor RGNF
PhD Jasna Orešković, Associate Professor RGNF
PhD Tomislav Korman, Associate Professor RGNF

Defence date: February 24, 2023., Faculty of Mining, Geology and Petroleum Engineering, University of Zagreb

Sadržaj

POPIS SLIKA

POPIS TABLICA

1. UVOD

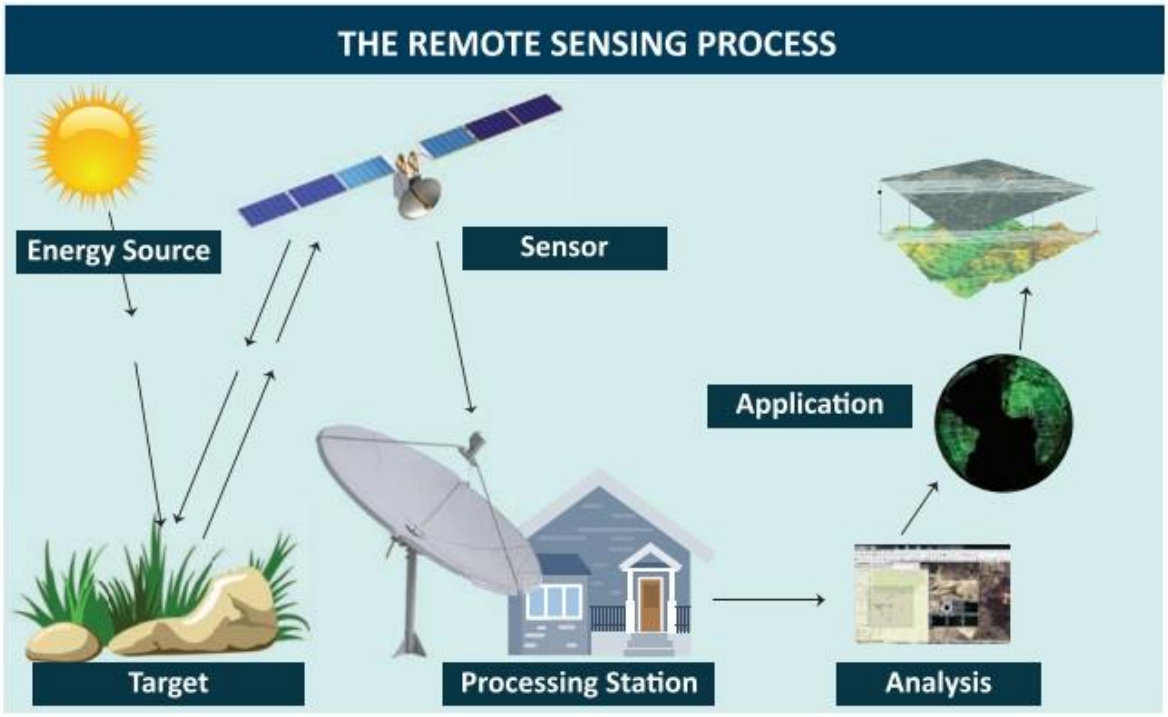
Daljinska istraživanja su naziv za proces opažanja i praćenja fizičkih karakteristika područja, mjerenjem njegovog emitiranog i reflektiranog zračenja (*USGS, 2023.)*. Istraživanja se provode pomoću satelitskih snimki, snimanjem iz aviona, pomoću dronova, brodova ili nekim drugim načinom koji nema direktan dodir sa istraživanim područjem. Korišteni su podaci SENTINEL-2 satelita koji je rezultat suradnje ESA-e (Europske svemirske agencije) i Europskog programa Copernicus.

Cilj ovog diplomskog rada je klasifikacija SENTINEL-2 satelitske snimke na području Grada Zagreba te usporedba rezultata algoritama korištenih za klasifikaciju. Korištena su tri algoritma od kojih je jedan za nenadziranu klasifikaciju (*K-means*) te dva za nadziranu klasifikaciju (*random forest i xtreme gradient boost*).

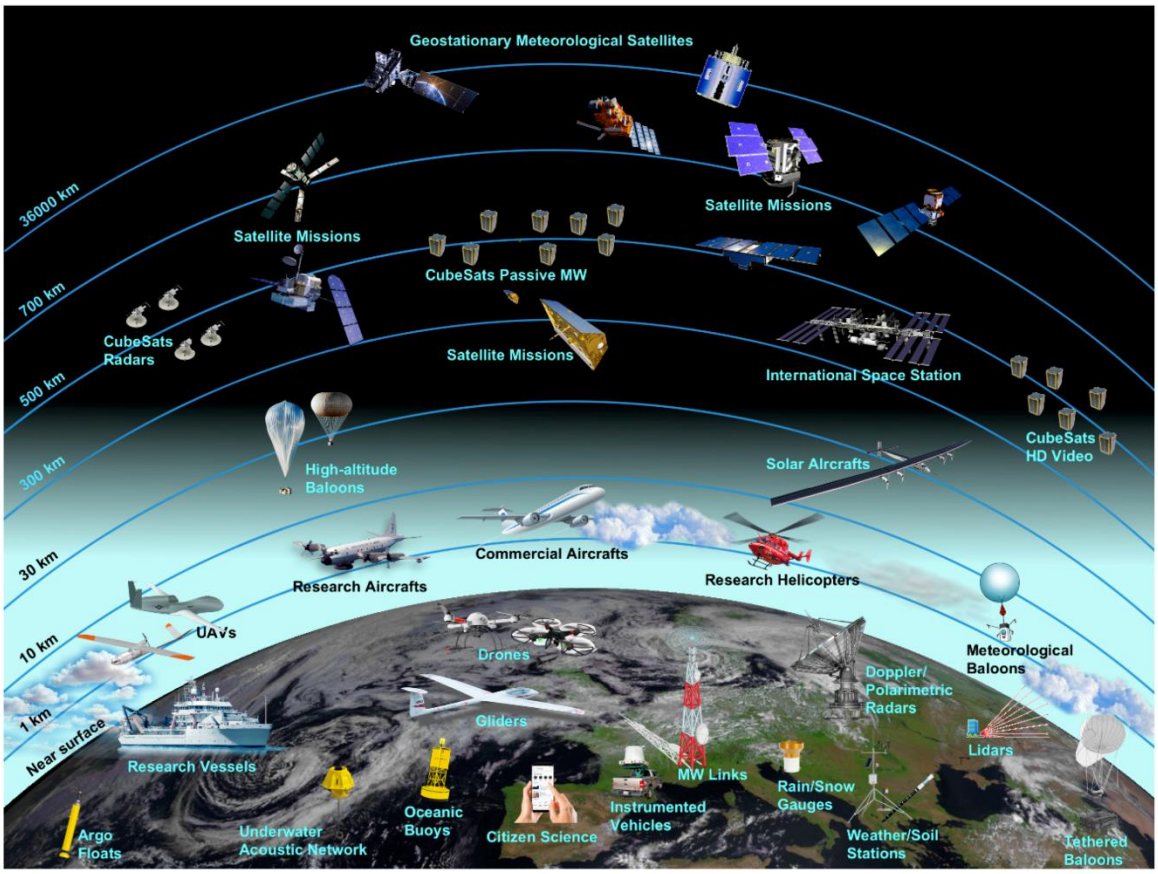
Prikupljanje podataka potrebnih za klasifikaciju odradilo se u softveru QGIS dok su se rezultati dobili u programskom jeziku R za koji je korišteno grafičko sučelje R-Studio. Dobivena su ukupno tri grafička prikaza (po jedan za svaku metodu) te grafički prikaz spektralnog profila. Provedena je statistička usporedba rezultata te je određen algoritam koji je pokazao najveću razinu točnosti.

2. DALJINSKA ISTRAŽIVANJA

Daljinsko istraživanje je proces otkrivanja i praćenja fizičkih karakteristika područja mjerenjem njegovog reflektiranog i emitiranog zračenja na daljinu, obično sa satelita ili zrakoplova (Slika 2.1 i Slika 2.2). Neki od načina korištenja daljinskih snimaka Zemlje uključuju: mapiranje velikih šumskih požara iz svemira, praćenje oblaka za pomoć u predviđanju vremena, promatranje erupcije vulkana, pomoć u promatranju pješčanih oluja, praćenje rasta grada i promjena u poljoprivrednom zemljištu ili šumskom području tijekom nekog vremenskog perioda (*USGS, 2023.*).



Slika 2.1 *Shematski prikaz daljinskih istraživanja (Brainkart, 2023.)*



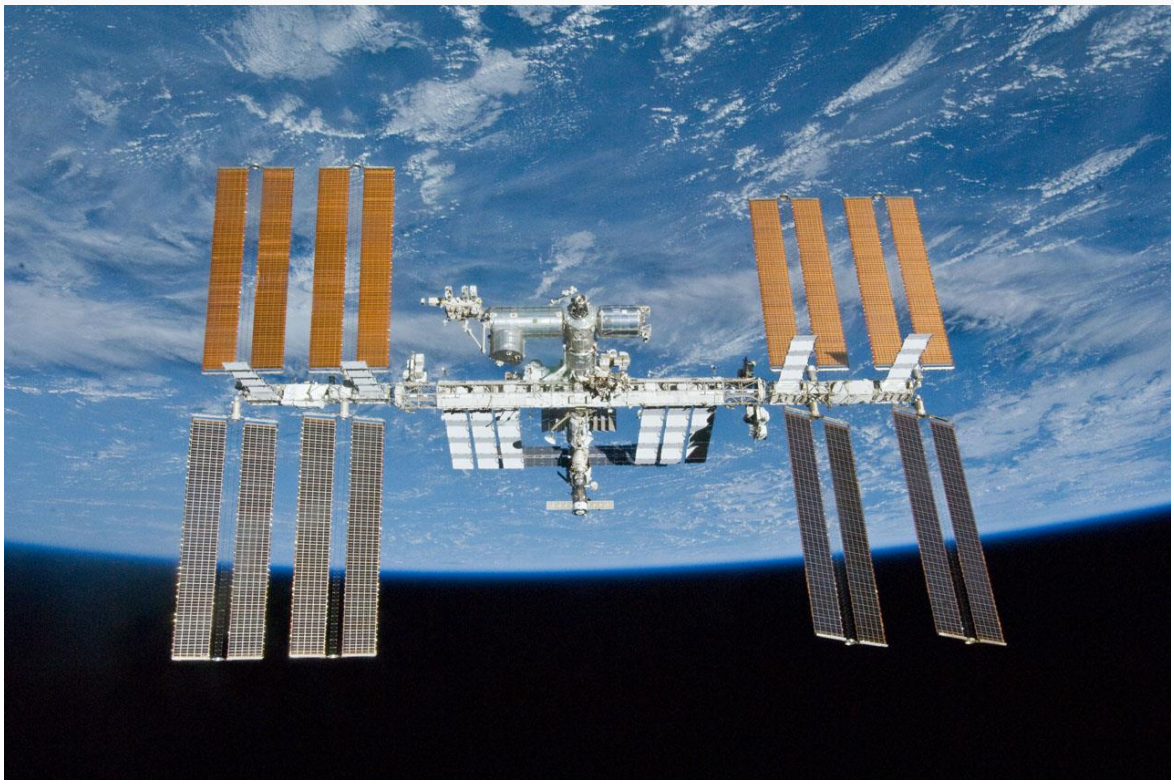
Slika 2.2 *Prikaz daljinskih istraživanja iz više izvora (Levizzani, V. i dr., 2019.)*

2.1 Sateliti

Satelit je tijelo koje kruži oko drugog tijela u svemiru. Postoje dvije različite vrste satelita: prirodni i umjetni. Primjeri prirodnih satelita su Zemlja i Mjesec. Zemlja se okreće oko Sunca, a Mjesec oko Zemlje. Umjetni satelit je objekt koji se lansira u svemir i kruži oko tijela u svemiru. Primjeri satelita koje je napravio čovjek uključuju svemirski teleskop Hubble i Međunarodnu svemirsku postaju koja je prikazana na slici (Slika 2.3).

Sateliti koje je izradio čovjek dolaze u raznim oblicima i veličinama te imaju različite instrumente na sebi za obavljanje različitih funkcija dok su u svemiru. Sateliti su napravljeni od strane inženjera i potrebni su mjeseci, ponekad čak i godine da se dovrši konstrukcija satelita. Sateliti moraju izdržati mnoge testove kako bi bili sigurni da mogu izdržati lansiranje i izazovno okruženje svemira.

NASA uspostavlja misije za određenu svrhu, a inženjeri razvijaju satelit za obavljanje potrebnih funkcija za tu misiju. Nakon što je satelit lansiran u svemir, Svemirske komunikacije i navigacija osiguravaju kanal komunikacije za podatke koji idu prema i od Zemlje i satelita. Ove komunikacije uključuju naredbe svemirskim letjelicama kao i znanstvene podatke koji dolaze na Zemlju *(NASA-a, 2014.)*.



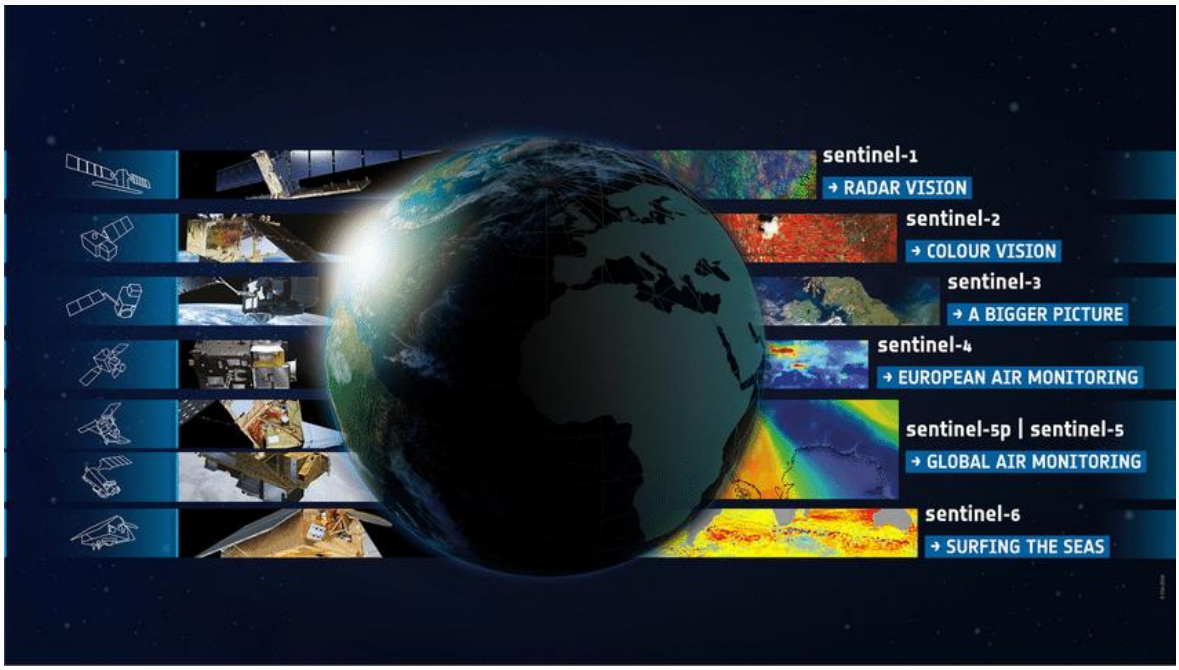
Slika 2.3 *Međunarodna svemirska postaja (NASA-b, 2020.)*

2.2 Sentinel serija satelita

ESA (Europska svemirska agencija) razvila je niz misija nove generacije za promatranje Zemlje u kojima sudjeluju ESA i Europski program Copernicus.

Copernicus je program Europske unije za promatranje Zemlje. Cilj programa je promatranje našeg planeta i njegovog okoliša u korist svih europskih građana. Nudi informacije koje se crpe iz satelitskog promatranja Zemlje i podataka in situ. Cilj programa Sentinel je zamijeniti starije misije za promatranje Zemlje koje su se povukle ili su trenutno pri kraju svog radnog vijeka, poput misija ERS (European remote sensing satellite) i Envisat (Environmental Satellite). Time će se osigurati kontinuitet podataka kako ne bi bilo praznina u studijama koje su u tijeku. Svaka se misija fokusira na drugačiji

aspekt promatranja Zemlje: praćenje atmosfere, oceana i kopna, a podaci su korisni u mnogim primjenama. Postoji 7 Sentinel misija. Sentinel – 1 s ciljem praćenja kopna i oceana, sastavljen od dva satelita koji rade danju i noću. Cilj Sentinela-2 je praćenje kopna, a misija se sastoji od dva satelita koji pružaju optičke slike visoke rezolucije. Primarni cilj Sentinela-3 je promatranje mora, a proučava topografiju površine mora, površinsku temperaturu mora i kopna, boju oceana i kopna te se sastoji od tri satelita. Sentinel-4 posvećen je praćenju kvalitete zraka. Cilj misije je osigurati kontinuirano praćenje sastava Zemljine atmosfere u visokoj vremenskoj i prostornoj rezoluciji. Sentinel-5 posvećen je praćenju kvalitete zraka. Pruža podatke o kvaliteti zraka na širokoj globalnoj razini. Preteča satelitske misije Sentinel-5 bila je misija Sentinel-5P, koja je imala cilj popuniti prazninu u podacima i osigurati kontinuitet podataka između povlačenja satelita Envisat i NASA-ine misije Aura te lansiranja Sentinela-5. Sentinel-6 je misija kojoj je cilj mjerenje visine površine mora, To je misija razvijena za praćenje srednje razine mora i stanja u oceanima. (*ESA, 2023.*). Slikovni pregled svih misija prikazan je na slici (Slika 2.4).



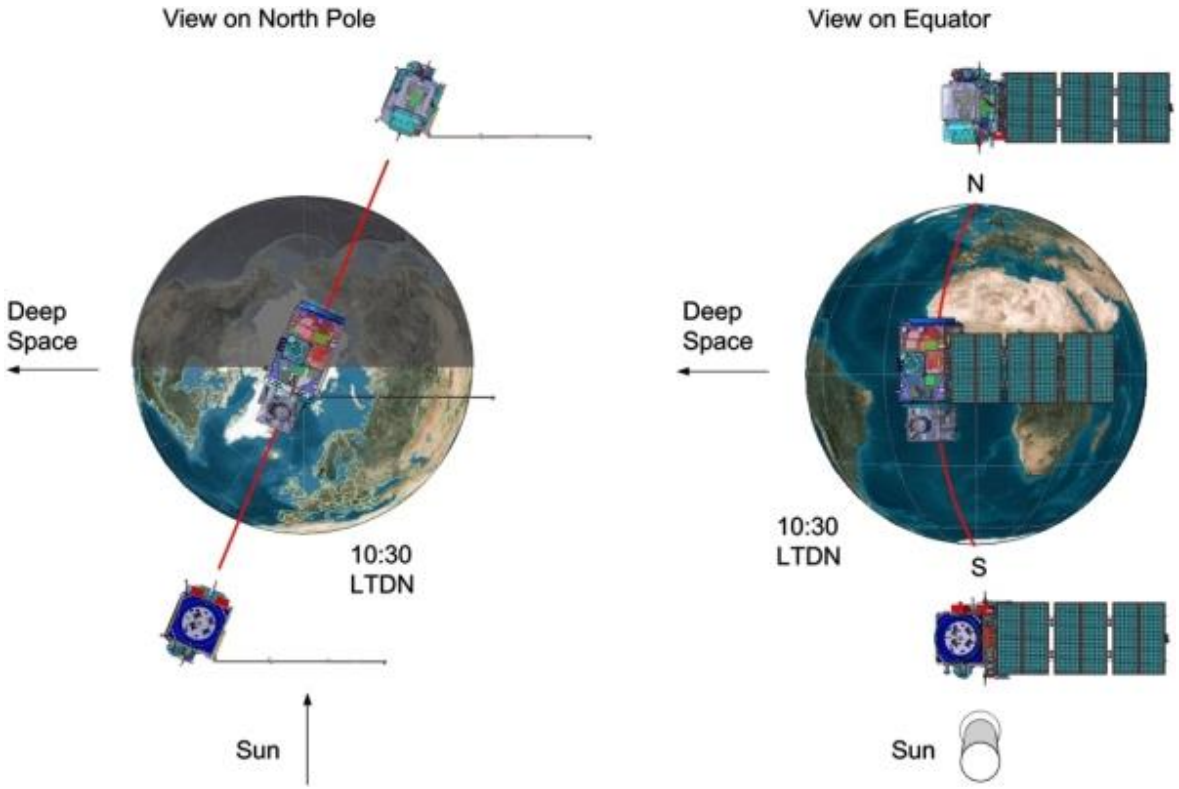
Slika 2.4 *Slikovni pregled Sentinel misija (ESA, 2023.)*

2.3 SENTINEL-2

Misija Copernicus Sentinel-2 sastoji se od dva satelita u polarnoj orbiti usmjerenih pod kutom od 180° jedan prema drugom što je prikazano na slici (Slika 2.6). Sentinel-2 koji je

prikazan na slici (Slika 2.5) nosi teret optičkog instrumenta koji uzorkuje 13 spektralnih pojasa: četiri pojasa na 10 m, šest pojaseva na 20 m i tri pojasa na 60 m prostorne rezolucije. Širina orbitalnog pojasa je 290 km. Ciljevi misije Sentinel-2 su osigurati: sustavno globalno prikupljanje multispektralnih snimaka visoke rezolucije povezanih s visokom učestalošću ponovnih posjeta, kontinuitet multispektralnih snimaka koje pruža SPOT serija satelita i USGS LANDSAT Thematic Mapper instrument, podatke o promatranju za sljedeću generaciju operativnih proizvoda, kao što su karte pokrova zemljišta, karte za otkrivanje promjena zemljišta i geofizičke varijable. Prikupljeni podaci, pokrivenost misije i velika učestalost ponovnih posjeta omogućavaju stvaranje geoinformacija na lokalnoj, regionalnoj, nacionalnoj i međunarodnoj razini. Podaci su osmišljeni tako da ih mogu mijenjati i prilagođavati korisnici zainteresirani za tematska područja kao što su: prostorno planiranje, agroekološki monitoring, praćenje vode, monitoring šuma i vegetacije, zemljišni ugljik, praćenje prirodnih resursa i globalno praćenje usjeva (*ESA, 2023.*).

Slika 2.5 *SENTINEL-2 satelit (Copernicus-a, 2023.)*



Slika 2.6 *Orbitalna konfiguracija dvostrukog satelita SENTINEL-2 (ESA, 2023.)*

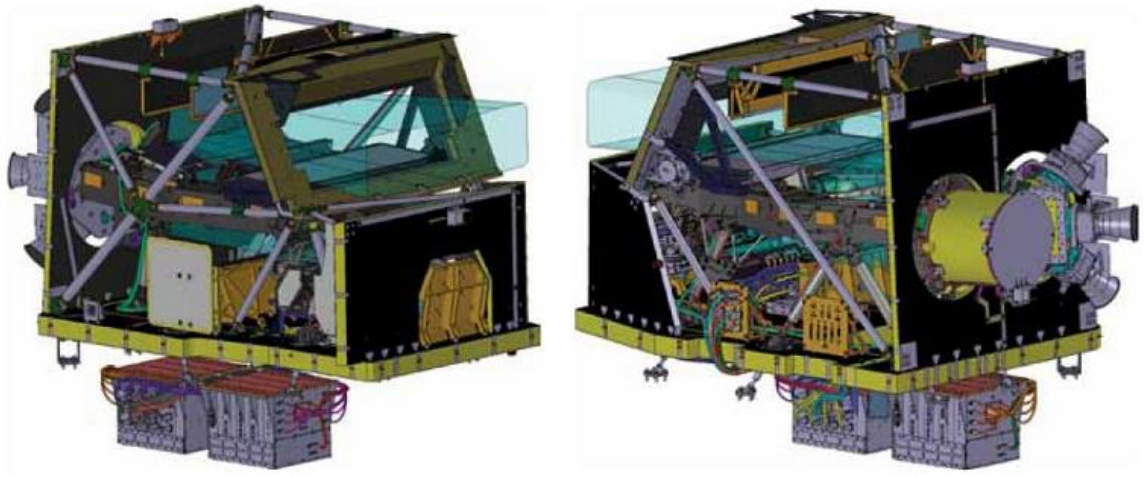
2.4 Multi spektralni instrument (MSI)

Dizajn multispektralnog instrumenta (MSI) ugrađenog u SENTINEL-2 vođen je zahtjevom za velikim opsegom, visokom geometrijskom i spektralnom izvedbom mjerenja.

MSI mjeri reflektirani sjaj Zemlje u 13 spektralnih pojaseva različitih valnih duljina i širina od VNIR (visible) do SWIR (short wave infrared) što je prikazano u tablici (Tablica 2-1). Izgled instrumenta prikazan je na slici (Slika 2.7).

Tablica 2-1 *Spektralni kanali, valna duljina, širina i rezolucija Sentinel 2 satelita (Copernicus-b, 2023.)*

	S2A		**S2B**		
Pojas	**Valna duljina (nm)**	**Širina (nm)**	**Valna duljina (nm)**	**Širina(nm)**	**Rezolucija (m)**
1	442.7	20	442.3	20	60
2	492.7	65	492.3	65	10
3	559.8	35	558.9	35	10
4	664.6	30	664.9	31	10
5	704.1	14	703.8	15	20
6	740.5	14	739.1	13	20
7	782.8	19	779.7	19	20
8	832.8	105	832.9	104	10
8a	864.7	21	864.0	21	20
9	945.1	19	943.2	20	60
10	1373.5	29	1376.9	29	60
11	1613.7	90	1610.4	94	20
12	2202.4	174	2185.7	184	20



Slika 2.7 *Multi spektralni instrument (ESA, 2023.)*

3. NDVI indeks

Indeks normalizirane razlike vegetacije (NDVI) normalizira reflektiranje vegetacije u bliskim infracrvenim valnim duljinama (NIR) uz apsorpciju klorofila u crvenim valnim duljinama (VNIR - RED). Dakle, to je kvantificiranje vegetacije mjerenjem razlike između bliskog infracrvenog (koje vegetacija snažno reflektira) i crvenog svjetla (koje vegetacija apsorbira).

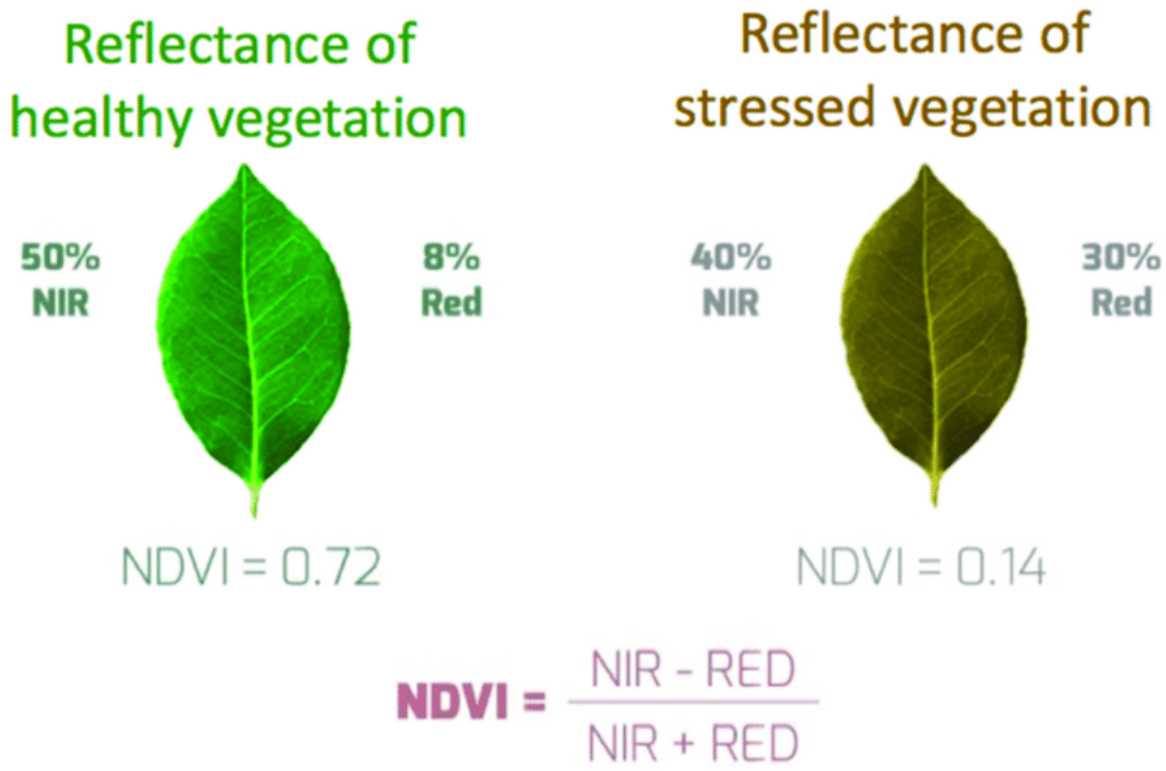
Raspon vrijednosti NDVI je od -1 do 1. Negativne vrijednosti NDVI (vrijednosti koje se približavaju -1) odgovaraju vodi. Vrijednosti blizu nule (-0,1 do 0,1) općenito odgovaraju neplodnim područjima kamenja, pijeska ili snijega. Niske, pozitivne vrijednosti predstavljaju grmlje i travnjake (približno 0,2 do 0,4), dok visoke vrijednosti označavaju umjerene i tropske prašume (vrijednosti koje se približavaju 1). Dobar je indikator za živu zelenu vegetaciju što je vidljivo na slici (Slika 3.1). (*Custom-scripts, 2023.*)

Indeks normalizirane razlike vegetacije, skraćeno NDVI, definiran je kao:

$$NDVI \coloneqq Indeks(NIR, RED) = \frac{NIR - RED}{NIR + RED} \qquad 3\text{-}1$$

Sukladno jednadžbi (3-1) za Sentinel-2 izračun NDVI definiran je kao:

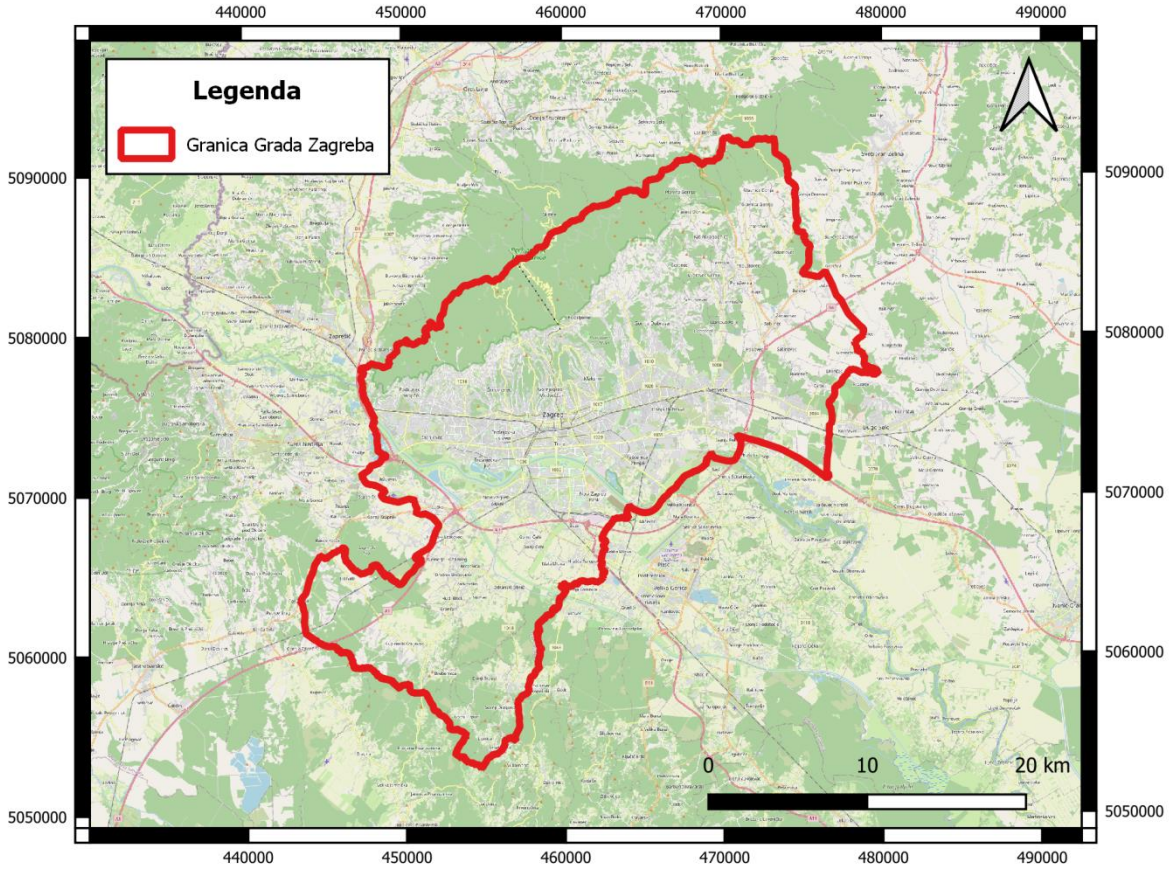
$$NDVI \coloneqq Indeks(Pojas\ 8, Pojas\ 4) = \frac{Pojas\ 8 - Pojas\ 4}{Pojas\ 8 + Pojas\ 4} \qquad 3\text{-}2$$



Slika 3.1 *Primjer NDVI-a za različita stanja vegetacije (Loures, L. i dr., 2020.)*

4. LOKACIJA PODRUČJA ISTRAŽIVANJA

Grad Zagreb je glavni grad Republike Hrvatske smješten u središnjoj Hrvatskoj i najveći grad po broju stanovnika (767,131 prema zadnjem popisu stanovništva 2021.). Grad Zagreb je vlastita administrativna jedinica ukupne površine 641 km^2, a tvore je uže gradsko područje i 68 naselja. Administrativna granica Grada Zagreba definirana je hrptom Medvednice na Sjeveru, središnji dio čine lijeva i desna obala rijeke Save, a južnu granicu predstavljaju Vukomeričke Gorice. Slika 4.1 prikazuje smještaj područja Grada Zagreba (*Halapir, I., 2022.*).



Slika 4.1 *Područje Grada Zagreba u softveru QGIS*

5. INFORMATIČKE OSNOVE RADA

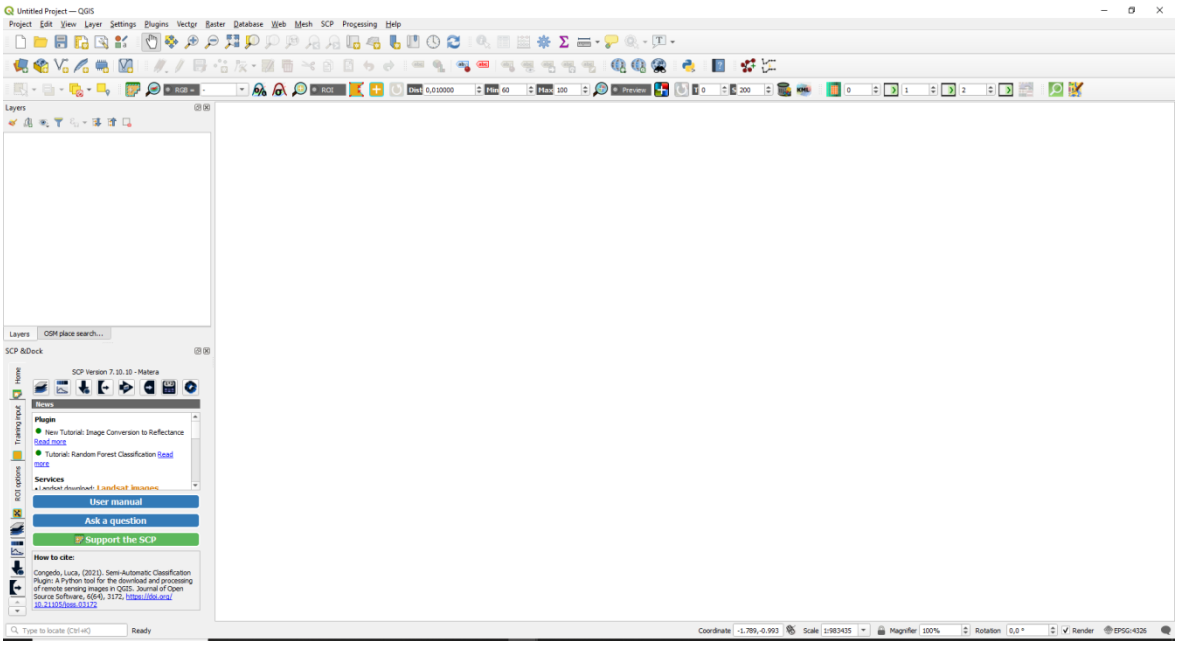
U ovom poglavlju bit će objašnjeni softveri koji su se koristili u procesu izrade ovog diplomskog rada.

5.1 GIS – Geografski informacijski sustav i QGIS

Geografski informacijski sustav (GIS) je alat za organiziranje, analizu i dijeljenje podataka temeljenih na lokaciji. GIS se može koristiti za upravljanje gotovo svim vrstama podataka: proračunskim tablicama, fotografijama, topografskim kartama, satelitskim i zračnim snimkama, podacima daljinskog senzora, nacrtima i planovima. Vizualizacija podataka pomoću GIS mapiranja omogućuje nam da pratimo napredak izgrađenih površina tijekom vremena, analiziranje prostornih odnosa, identificiranje područja rizika, iscrtavanje rute prijevoza i uspoređivanje prirodnih značajki s ljudskim djelovanjem (*Unearth labs, 2023.).*

Podaci u GIS-u mogu se podijeliti u dvije kategorije: prostorno referencirani podaci koji mogu biti vektorski i rasterski te atributne tablice koje se prikazuju u tabelarnom formatu (*GIS Lounge, 2022.*).

QGIS (ranije Quantum GIS) je geografski informacijski sustav otvorenog koda (GIS). Ovaj softver je besplatna alternativa vlasničkom GIS softveru kao što su ESRI ArcGIS proizvodi koji mogu biti vrlo skupi. QGIS uključuje slične funkcije i značajke kao ArcGIS i omogućuje korisnicima prikaz, manipuliranje i stvaranje prostornih podataka. Podržava različite ekstenzije datoteka prostornih podataka (.shp, .tif, .csv, .img, itd.) i kompatibilan je s operativnim sustavima Linux, Unix, Mac i Windows (*Iowa State University, 2014.*). Slika 5.1 prikazuje QGIS sučelje.



Slika 5.1 *Prikaz QGIS sučelja prije učitavanja podataka*

5.2 R softver i R Studio

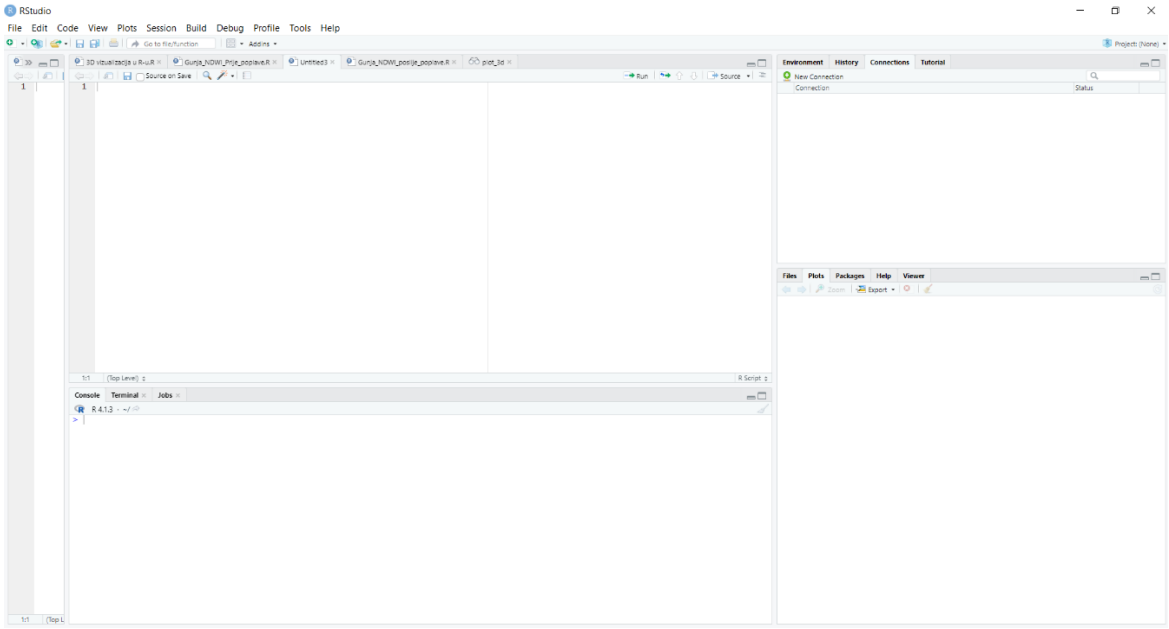
R je programski jezik i softversko okruženje za statističku analizu, grafički prikaz i izvještavanje. Besplatno je dostupan pod GNU General Public License, a unaprijed kompajlirane binarne verzije dostupne su za različite operativne sustave kao što su Linux, Windows i Mac. Ovaj programski jezik nazvan je R prema prvom slovu imena dvojice autora R-a (Robert Gentleman i Ross Ihaka) (*Tutorialspoint, 2023.*).

R ne samo da nam omogućuje grananje i petlje, već također omogućuje modularno programiranje pomoću funkcija. R omogućuje integraciju s procedurama napisanim u jezicima C, C++, .Net, Python i FORTRAN radi poboljšanja učinkovitosti.

U današnje doba, R je jedan od najvažnijih alata koji koriste istraživači, analitičari podataka, statističari i trgovci za dohvaćanje, čišćenje, analizu, vizualizaciju i prezentaciju podataka (*Javapoint-a, 2023.*).

R Studio je integrirano razvojno okruženje (IDE) odnosno skup integriranih alata dizajniranih da povećaju produktivnost korisnika s R programskim jezikom. Uključuje konzolu, uređivač za označavanje sintakse koji podržava izravno izvršavanje koda i niz

robusnih alata za iscrtavanje, pregled povijesti, otklanjanje pogrešaka i upravljanje radnim prostorom (*RStudio, 2023.).* Izgled sučelja R Studio-a prikazan je na slici (Slika 5.2).



Slika 5.2 *R Studio sučelje*

5.3 R „Paketi“

R paketi su zbirke funkcija i skupova podataka koje je razvila zajednica. Oni povećavaju mogućnosti R-a poboljšanjem postojećih osnovnih funkcionalnosti ili dodavanjem novih. Paket obično uključuje kod, dokumentaciju za paket i funkcije unutar njega, neke testove za provjeru funkcioniranja i skupove podataka (*Datacamp, 2019.*).

Paketi koji su korišteni u ovom diplomskom radu prikazani su i opisani u tablici (Tablica 5-1).

Tablica 5-1 *Tablica korištenih paketa*

Naziv R paketa	Svrha paketa
Raster	Paket implementira osnovne funkcije i funkcije visoke razine za rasterske podatke i operacije s vektorskim podacima.
Terra	Omogućuje analizu prostornih podataka s

	rasterskim i vektorskim podacima. Uključuje metode predviđanja i interpolacije koje olakšavaju upotrebu modela regresijskog tipa za prostorno predviđanje, uključujući podatke satelitskog daljinskog istraživanja.
xgboost	Paket uključuje učinkovit alat za rješavanje linearnih modela i algoritme učenja stabla. Paket može automatski raditi paralelno izračunavanje na jednom stroju što bi moglo biti više od 10 puta brže od postojećih paketa za povećanje gradijenta. Podržava različite objektivne funkcije, uključujući regresiju, klasifikaciju i rangiranje. Paket je napravljen da bude proširiv, tako da je korisnicima omogućeno i jednostavno definiranje vlastitih ciljeva.
rgdal	Omogućuje vezanje na 'Geospatial' Data Abstraction Library ('GDAL') (>= 1.6.3) i pristup operacijama projekcije/transformacije iz 'PROJ.4' biblioteke. I 'GDAL' rasterski i 'OGR' podaci vektorske karte mogu se uvesti u R, a 'GDAL' rasterski podaci i 'OGR' vektorski podaci eksportirati. Koriste se klase definirane u paketu 'sp'.
ExtractTrainData	Korištenjem multispektralne slike i ESRI datoteke oblika (točka/linija/poligon), generira se podatkovna tablica za klasifikaciju, regresiju ili drugu obradu. Tablica podataka sadržavat će rasterske

	vrijednosti po pojasu i ID-ove datoteke oblika (korisnički definirane).
randomForest	Klasifikacija i regresija temeljena na šumi drveća korištenjem slučajnih inputa.
caret	Caret paket (skraćenica od Classification And REgression Training) skup je funkcija koje pokušavaju pojednostaviti proces stvaranja prediktivnih modela.
Metrics	Implementacija metrike procjene u R-u koja se obično koristi u nadziranom strojnom učenju. Implementira metriku za regresiju, vremenske serije, binarnu klasifikaciju, klasifikaciju i probleme pronalaženja informacija.
e1071	Funkcije za analizu latentne klase, kratku Fourierovu transformaciju, neizrazito klasteriranje, strojeve potpornih vektora, izračunavanje najkraćeg puta, klasteriranje u vrećicama, Bayesov klasifikator, generalizirani k-najbliži susjed ...

5.4 Instalacija „paketa“

Instalacija R „paketa“ izvršena je u samom R Studio-u. Instalacija u R Studio-u izvodi se funkcijom *'install.packages(„naziv paketa“)*. Nakon instalacije u određeni direktorij paket je spreman za uporabu, ali ga je potrebno ručno učitati u konzoli naredbom *library(„naziv paketa“)*.

6. TEORETSKE OSNOVE KLASIFIKACIJSKIH ALGORITAMA

6.1 Nenadzirana klasifikacija (K-means)

K-Means je klastering (eng. *clustering*) algoritam nenadziranog učenja, koji grupira neoznačeni skup podataka u različite klastere. Ovdje K definira broj unaprijed definiranih klastera koji se trebaju stvoriti u procesu. To je iterativni algoritam koji dijeli neoznačeni skup podataka u k različitih klastera na takav način da svaki skup podataka pripada samo jednoj grupi koja ima slična svojstva. Taj princip prikazan je na slici (Slika 6.1). S obzirom da je prigodan za neoznačene skupove podataka u kontekstu satelitskih snimaka praktičan je kada imamo snimku područja koje ne poznajemo te nam treba relativno brz rezultat. Omogućuje nam da neoznačene podatke grupiramo u više skupina odnosno klastera. To je algoritam temeljen na centroidu, gdje je svaki klaster pridružen centroidu. Glavni cilj ovog algoritma je minimizirati zbroj udaljenosti između podatkovne točke i njihovih odgovarajućih klastera (*Javapoint-b, 2023.*).

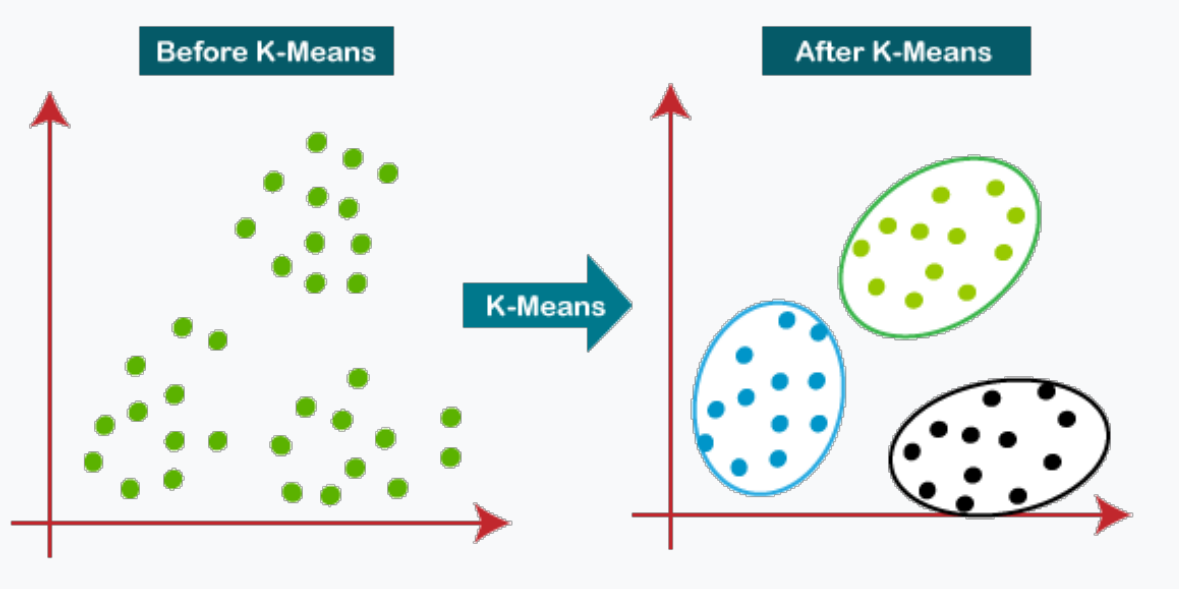
Postoji više algoritama unutar samog *K-means*, jedan o najpoznatijih je Llyod algoritam. Lloydov algoritam je iterativna metoda koja se koristi za pronalaženje optimalnog rješenja za problem klasteriranja *K-means* vrijednosti.

Algoritam ima dva koraka:

Korak dodjele: U ovom koraku svaka podatkovna točka se dodjeljuje najbližem centroidu, na temelju euklidske udaljenosti između podatkovne točke i centroida.

Korak ažuriranja: U ovom koraku centroidi se ažuriraju na srednju vrijednost svih podatkovnih točaka koje su im dodijeljene u prethodnom koraku.

Taj se proces ponavlja dok se centroidi ne konvergiraju, što znači da se dodjela podatkovnih točaka centroidima više ne mijenja. Konačni centroidi se smatraju optimalnim rješenjem za problem klasteriranja *K-means* (*Datascience lab, 2013.*).

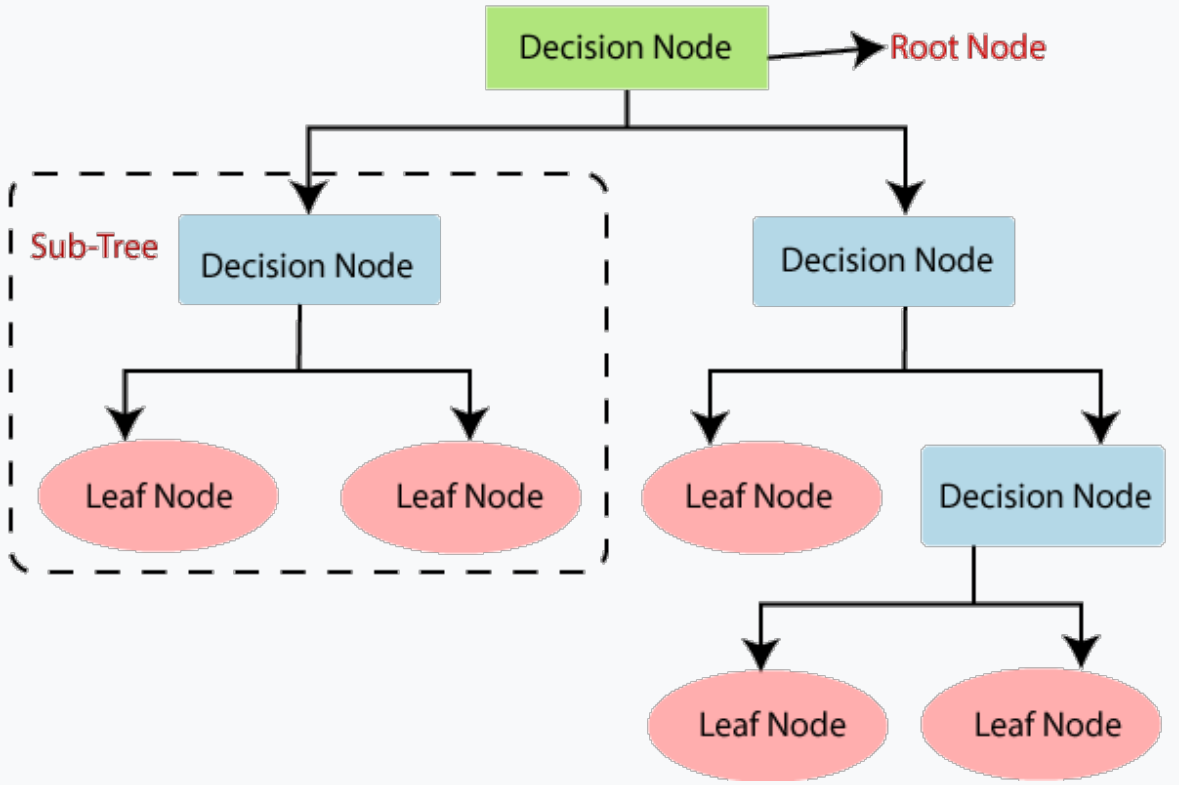


Slika 6.1 *Prikaz K-means algoritma (Javapoint-b, 2023.)*

6.2 Nadzirana klasifikacija

6.2.1 Decision tree algoritam

Stablo odlučivanja (eng. *decision tree*) je neparametarski nadzirani algoritam učenja koji se koristi za zadatke klasifikacije i regresije. Ima hijerarhijsku strukturu stabla koja se sastoji od korijenskog čvora (eng. *root node*), grana (eng. *branches*), čvorova odlučivanja ili unutarnjih čvorova (eng. *decision nodes*) i lisnih čvorova ili terminalnih (eng. *leaf nodes*). *Decision tree* počinje s korijenskim čvorom koji nema dolazne grane te u njega ulazi cijeli početni skup podataka. Odlazne grane iz korijenskog čvora zatim ulaze u čvorove odlučivanja. Na temelju dostupnih značajki, obje vrste čvorova provode procjene kako bi formirale homogene podskupove, koji su označeni lisnim čvorovima. Listni čvorovi predstavljaju sve moguće ishode unutar skupa podataka (*IBM, 2023.)*. Shematski prikaz *decision tree* modela prikazan je na slici (Slika 6.2).



Slika 6.2 *Shematski prikaz decision tree modela* (*Javapoint-c, 2023.*)

Za razumijevanje principa na kojima korijenski čvor i čvorovi odlučivanja formiraju uvjete za daljnju klasifikaciju podataka potrebno je objasniti dva ključna koncepta: entropija i informacijski dobitak.

Entropija je koncept koji proizlazi iz teorije informacija, koja mjeri nečistoću vrijednosti uzorka. Definirana je sljedećom jednadžbom (6-1):

$$Entropija\ (S) = -\sum_{c \epsilon C} p(c) log_2\, p(c) \qquad 6\text{-}1$$

Gdje:

S- predstavlja skup podataka za koji se izračunava entropija

C - predstavlja klase u skupu S

p(c) - predstavlja udio podatkovnih točaka koje pripadaju klasi c prema ukupnom broju podatkovnih točaka u skupu S

Vrijednosti entropije mogu biti između 0 i 1. Ako svi uzorci u skupu podataka S, pripadaju jednoj klasi, tada će entropija biti jednaka nuli. Ako je polovica uzoraka klasificirana kao jedna klasa, a druga polovica je u drugoj klasi, entropija će biti najveća, odnosno 1. Kako bi se odabrao najbolji uvjet za podjelu i pronašao optimalan *decision tree*, atribut s najmanjom vrijednosti entropije treba biti korišten.

Informacijski dobitak predstavlja razliku u entropiji prije i nakon podjele na danom atributu. Atribut s najvećim dobitkom informacija proizvest će najbolju podjelu jer daje najbolji rezultat u klasificiranju podataka. Dobitak informacija obično se predstavlja sljedećom formulom:

$$Informacijski\ dobitak\ (S, \alpha) = Entropija\ (S) \ - \sum_{v \in vcalues(\alpha)} \frac{|Sv|}{|S|}\ Entropija\ (Sv) \qquad 6\text{-}2$$

Gdje:

α-predstavlja određeni atribut ili oznaku klase

Entropija (S) - entropija skupa podataka

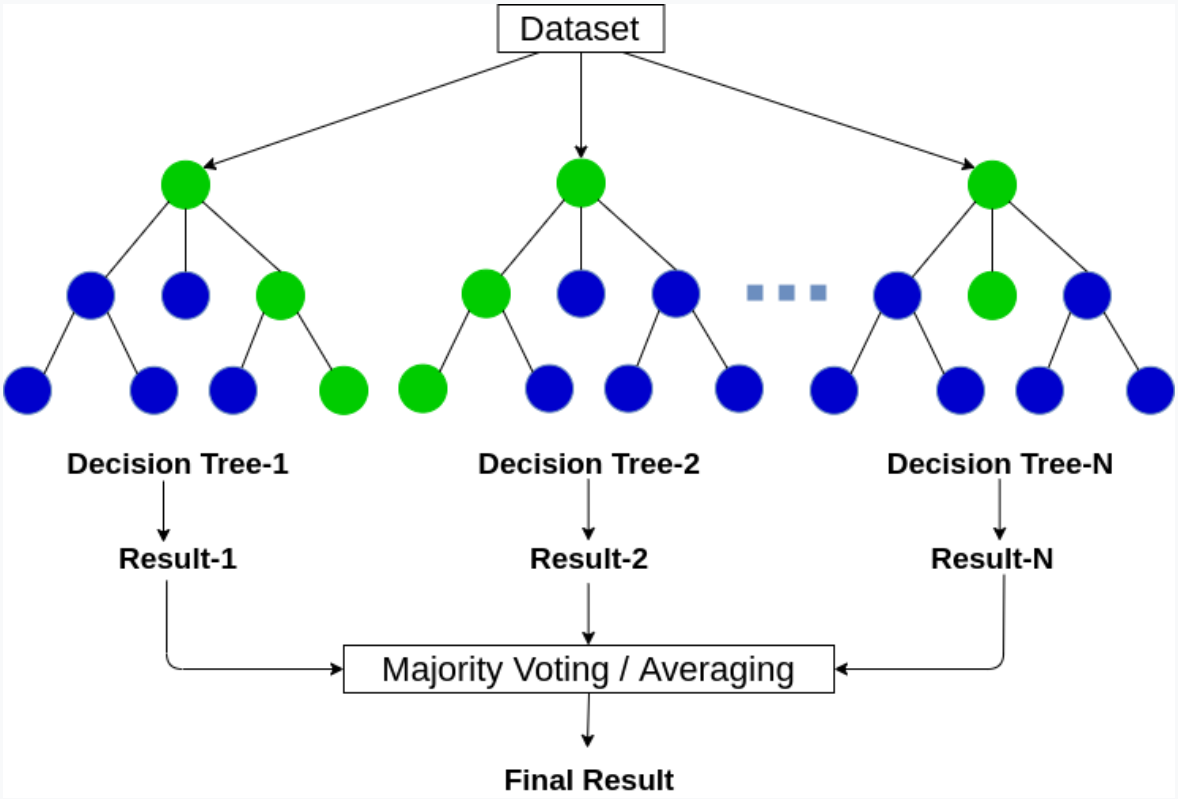
$\frac{|Sv|}{|S|}$ – omjer vrijednosti u skupu S_v prema broju vrijednosti u ukupnom skupu S

Entropija (S) - entropija skupa podataka S

6.2.2 Random forest algoritam

Random forest je algoritam koji se pokazao jako korisnim u klasifikacijskim problemima. Za razliku od osnovnog decision tree algoritma on je manje osjetljiv na promjenu u originalnim trening podacima zbog toga što uzima u obzir više nasumično stvorenih *decision trees*, potom podatak koji želimo klasificirati provodi kroz sva *decision trees* koja su stvorena te se klasa određuje „glasom većine“ s obzirom na ishod klasifikacije svakog zasebnog *decision tree-a*. Opisani princip modela grafički je prikazan na slici (Slika 6.3). Taj proces korištenja rezultata iz više zasebnih modela naziva se agregacija. Nad podacima se vrši nasumično uzorkovanje sa zamjenom odnosno „*bootstrapping*“. Algoritam je također poznat kao *bagging* algoritam te taj naziv proizlazi iz kombinacije riječi *boostrapping* i *agreggation*. *Bootstrapping* omogućuje modelu da bude manje osjetljiv na originalne trening podatke. Taj princip rada agregacijom predikcija različitih *decision trees* smanjuje varijancu (stupanj do kojeg predviđanja modela variraju ili fluktuiraju za različite

uzorke trening podataka). Model visoke varijance ima veliki raspon u svojim predviđanjima i može se previše prilagoditi (eng. *overfitting*) podacima za trening, dok model niske varijance ima mali raspon u svojim predviđanjima i vjerojatnije je da će se dobro generalizirati na nove podatke (*BMC, 2021.*).

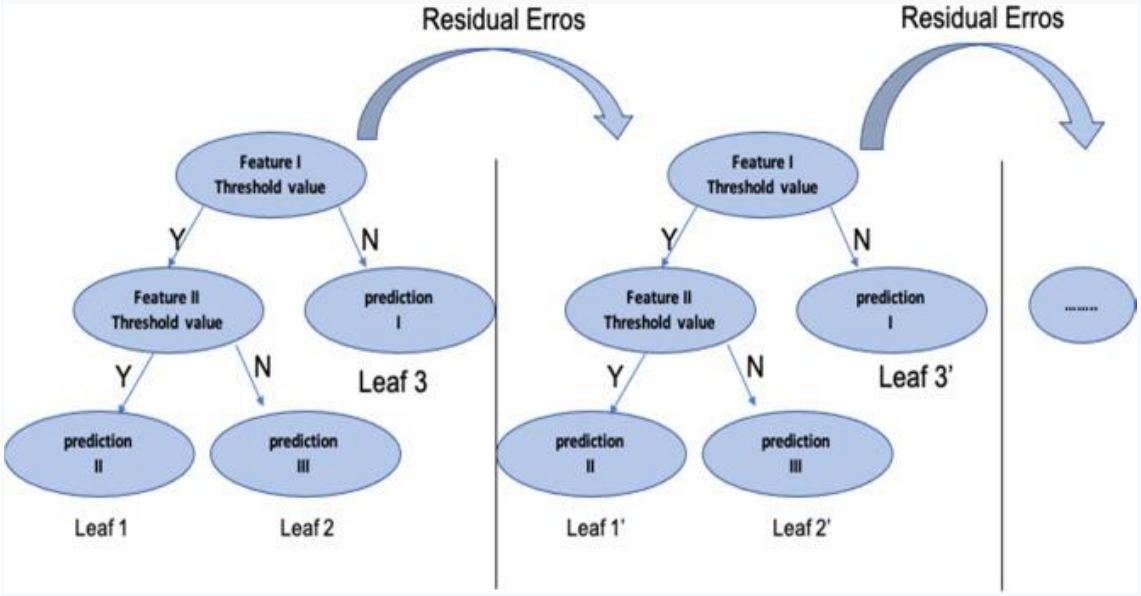


Slika 6.3 *Shematski prikaz random forest algoritma (AI Pool, 2021.)*

6.2.3 XGBoost (extreme gradient boost) algoritam

Boosting je u strojnom učenju metoda za stvaranje ansambla. Započinje prilagođavanjem početnog modela (npr. decision tree ili linearne regresije) podacima. Zatim se izrađuje drugi model koji se fokusira na točno predviđanje slučajeva u kojima prvi model ima lošu izvedbu. Očekuje se da će kombinacija ova dva modela biti bolja od svakog pojedinačnog modela. Zatim se ponavlja ovaj postupak pojačanja mnogo puta. Svaki sljedeći model pokušava ispraviti nedostatke kombiniranog pojačanog ansambla svih prethodnih modela. Oslanja se na intuiciju da najbolji mogući sljedeći model, u kombinaciji s prethodnim modelima, smanjuje ukupnu pogrešku predviđanja. Ključna ideja je postaviti ciljne ishode za ovaj sljedeći model kako bi se pogreška svela na minimum. Ciljani ishod za svaki slučaj u podacima ovisi o tome koliko promjena predviđanja tog slučaja utječe na ukupnu

pogrešku predviđanja. Ako mala promjena u predviđanju za slučaj uzrokuje veliki pad pogreške, tada je sljedeći ciljni ishod slučaja visoka vrijednost. Predviđanja iz novog modela koja su blizu njegovih ciljeva smanjit će pogrešku. Ako mala promjena u predviđanju za slučaj ne uzrokuje promjenu pogreške, tada je sljedeći ciljni ishod slučaja nula. Promjena ovog predviđanja ne smanjuje pogrešku. Naziv povećanje gradijenta proizlazi iz razloga što su ciljni ishodi za svaki slučaj postavljeni na temelju gradijenta pogreške u odnosu na predviđanje. Svaki novi model poduzima korak u smjeru koji minimizira pogrešku predviđanja, u prostoru mogućih predviđanja za svaki slučaj treninga (*Displayr, 2023.*). Na slici (Slika 6.4) prikazan je shematski prikaz prethodno opisanog boosting algoritma.

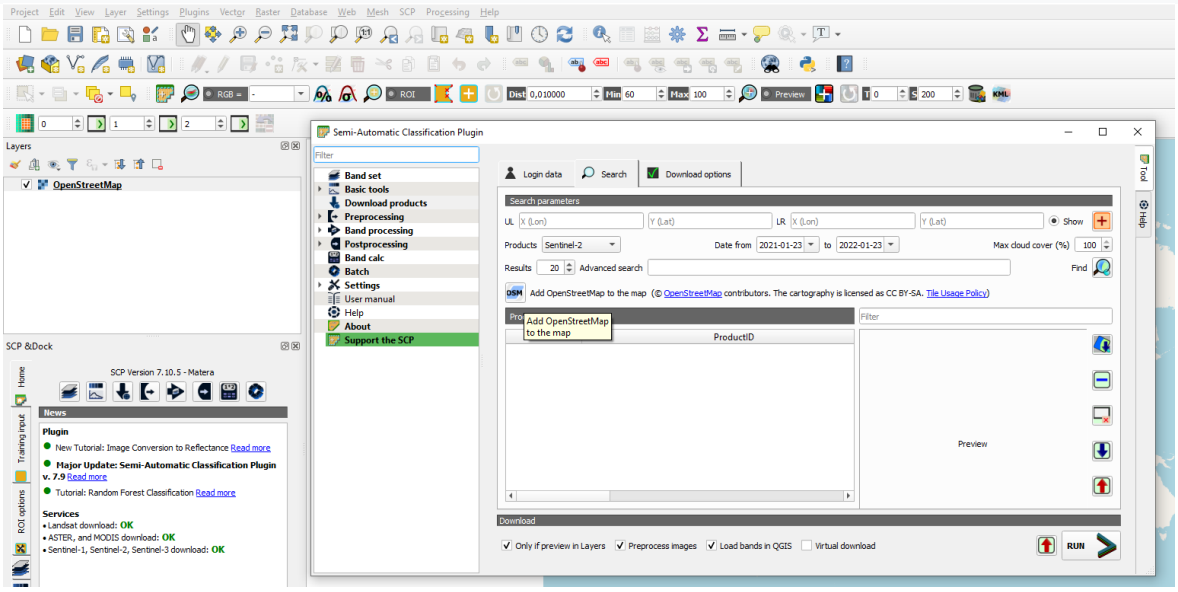


Slika 6.4 *Shematski prikaz gradient boosting algoritma (Science Direct, 2021.)*

7. PRIKUPLJANJE I OBRADA PODATAKA

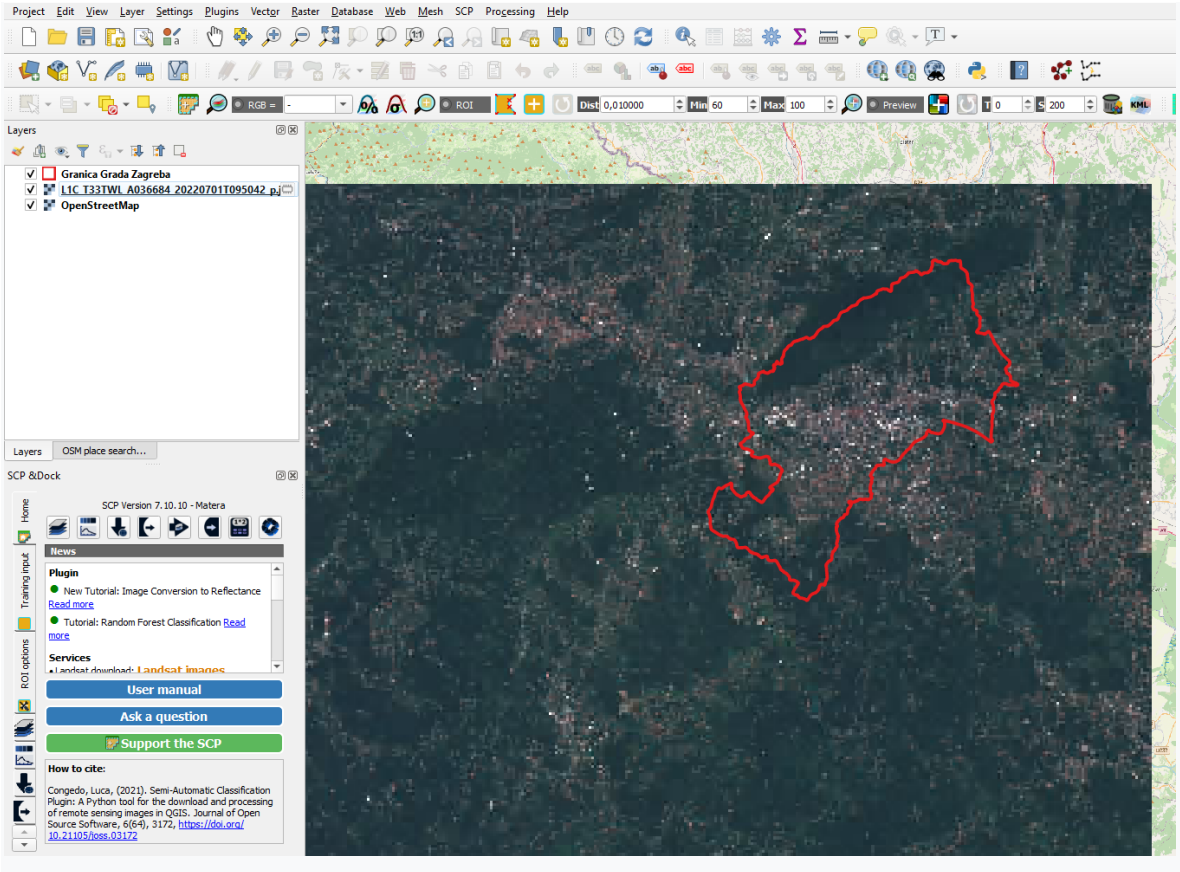
7.1 SENTINEL 2 satelitska snimka

Ulazni podatak za ovaj diplomski rad bila je Sentinel-2 satelitska snimka. Satelitska snimka preuzeta je u softveru QGIS pomoću SCP (*Semi-automatic-classification*) plugin-a (Slika 7.1).



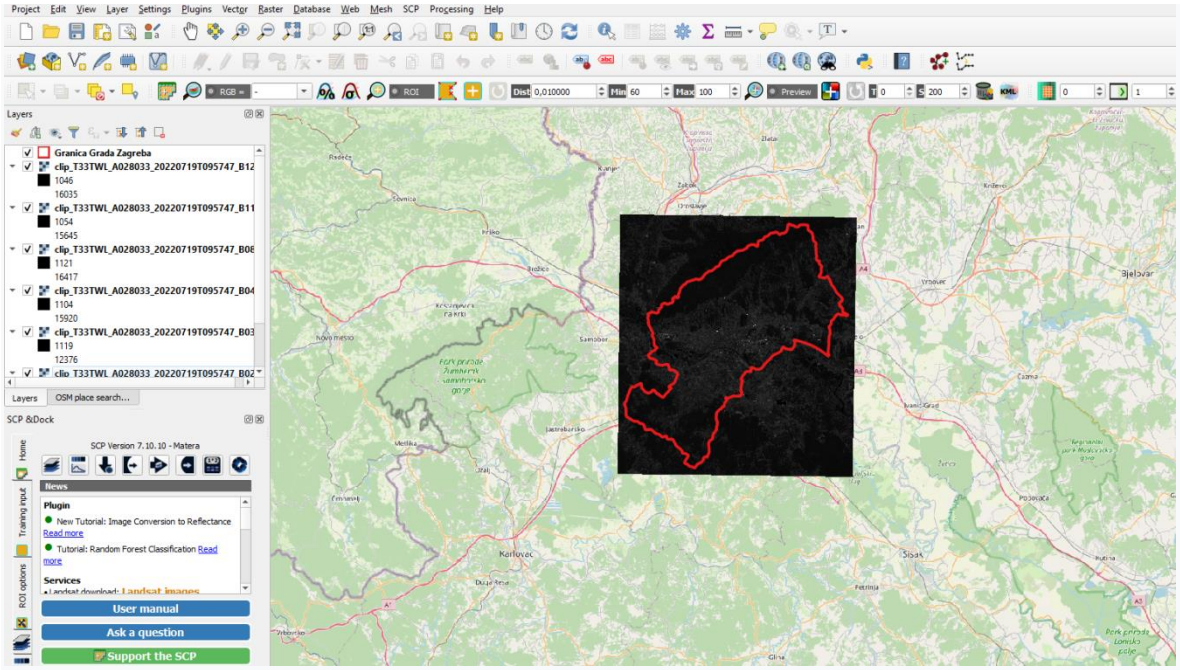
Slika 7.1 *Prikaz SCP plug-ina u softveru QGIS*

Učitan je OSM (*OpenStreetMap*) i područje Grada Zagreba u ESRI-jevom formatu za prostorne podatke (tzv. *shapefile*). S obzirom na poznatu granicu Grada Zagreba označeno je područje za koje se želi preuzeti satelitska snimka. Kako je za klasifikaciju potrebna snimka sa što manjom pokrivenošću oblacima (eng. *cloud cover*) postavljen je filter na 2% vrijednosti *cloud cover-a* i odabran je vremenski period od 1. kolovoza do 10. kolovoza 2022. s obzirom da je ljeti češća manja naoblaka. Izabrana je snimka sa malim iznosom *cloud cover-a* koji iznosi: 0.0014. Prikaz izabrane snimke prije preuzimanja željenih pojaseva prikazan je na slici (Slika 7.2).



Slika 7.2 *Prikaz izabrane satelitske snimke i granica Grada Zagreba*

Prije preuzimanja snimke odabrani su željeni pojasevi koji se žele preuzeti. Odabrani su pojasevi sa najpovoljnijom prostornom rezolucijom: pojas 2 (10m), pojas 3 (10m), pojas 4 (10m), pojas 8 (10m), pojas 11 (20m), pojas 12 (20m). Navedeni pojasevi će se u ovom radu unutar R koda smatrati pojasevima 1, 2, 3, 4 i 5. Kako bi algoritmi što brže radili sa rasterom, poželjna je što manja veličina snimke stoga je satelitska snimka izrezana na područje s koordinatama 442334.2652, 481119.3355, 5093835.2024, 5051429.1545 (HTRS96) što približno odgovara koordinatama (xmax, xmin, ymax, ymin) granice Grada Zagreba. Prikaz izrezanih pojaseva satelitske snime prikazan je na slici (Slika 7.3).



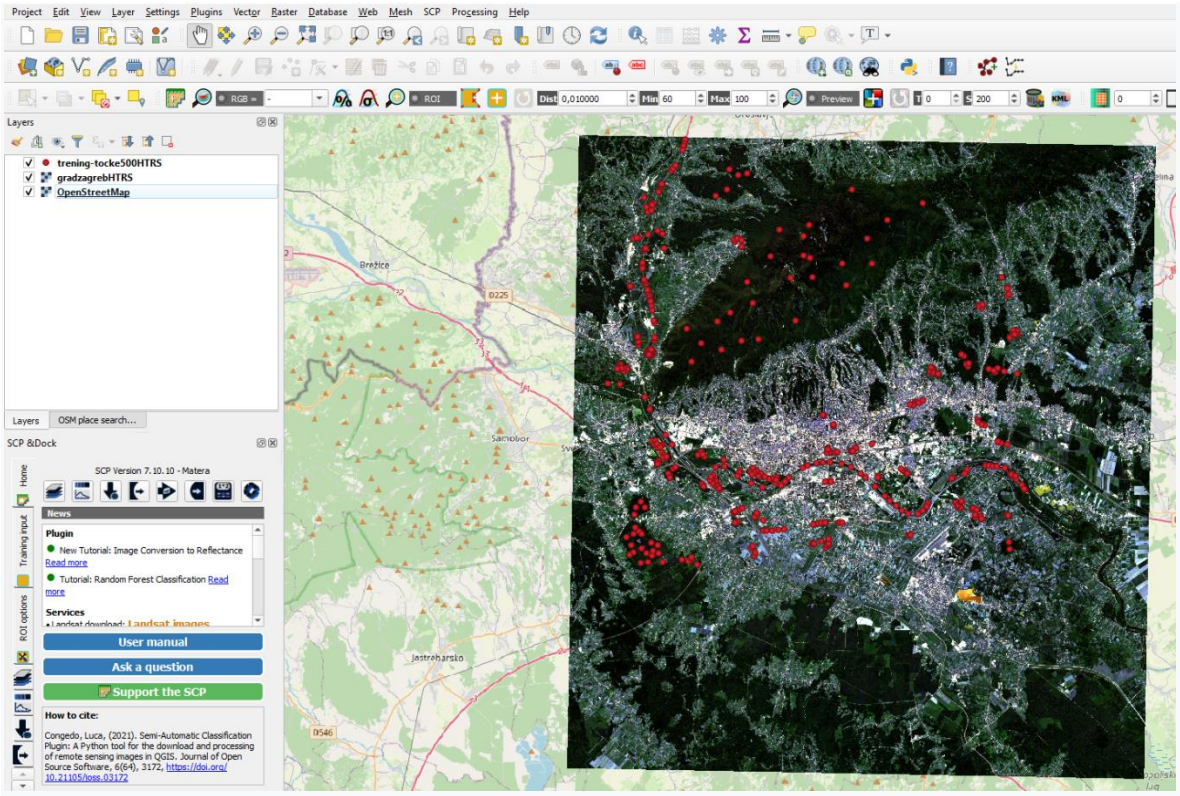
Slika 7.3 *Izrezani pojasevi prije stvaranja stack file-a*

S obzirom da se unutar programskog jezika R može manipulirati sa višeslojnim (eng. *multilayer*) podacima, izrezani rasteri svih 6 pojaseva potom su pretvoreni u jedan pomoću naredbe stack multiple rasters koja je prisutna u SCP pluginu te je izvezen *file* kao *GeoTIFF* sa 6 pojaseva koji će biti korišten u daljnoj obradi. To je napravljeno radi lakše daljnje provedbe koda te kako bi se izbjegla moguća ne slaganja u prostornom opsegu (eng. *extent)* između 10 i 20 metarskih pojaseva prilikom preklapanja slojeva u programskom jeziku R.

7.2 Trening podaci za nadziranu klasifikaciju

Za nadziranu klasifikaciju određeno je 500 trening podataka točkastog tipa. Određeno je pet različitih klasa: voda, šumska površina, izgrađeno područje, ceste, poljoprivredne površine.

U QGIS-u su numerirane brojevima od 1-5 zbog brzine stvaranja trening podataka te je veza između numeričkih oznaka i teksta sljedeća: 1- šumska površina, 2 – voda, 3 – izgrađena površina, 4 – poljoprivredna površina i 5 – cesta. Lokacija trening podataka prikazana je na slici (Slika 7.4).

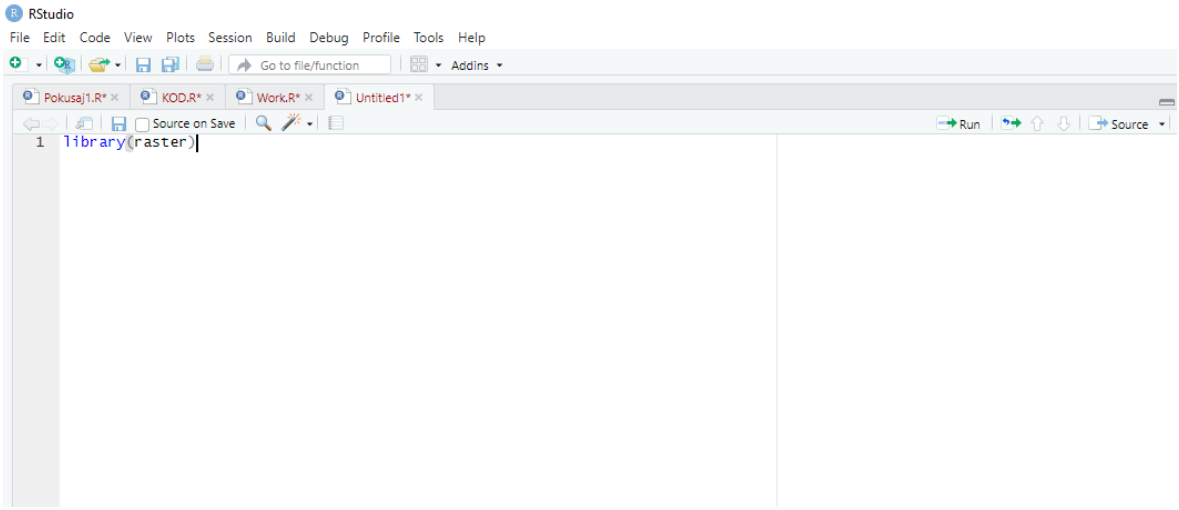


Slika 7.4 *Prikaz lokacije točkastih trening podataka (QGIS)*

Nakon stvaranja *shapefile-a* granica Grada Zagreba, trening podataka i rastera sa izrezanom satelitskom snimkom moguće je započeti sa klasifikacijom u softverskom jeziku R. Važno je napomenuti kako su sva tri file-a izvezena u istom koordinatnom sustavu (HTRS96) s obzirom da bi u slučaju ne poklapanja koordinatnog sustava između file-ova moglo doći će do grešaka u daljnoj obradi podataka u programskom jeziku R.

7.3 Učitavanje paketa, satelitske snimke i trening podataka u R-Studio-u

Prethodno instalirane pakete u R-u moguće je učitati naredbom *library („naziv paketa“)*. Primjer izgleda sučelja u R-studio-u tijekom učitavanja jednog od paketa vidi se na slici (Slika 7.5). Popis učitanih paketa sa detaljnim opisom vidljiv je u tablici (Tablica 2-1).



Slika 7.5 *Učitavanje paketa u programskom jeziku R*

Učitavanje satelitske snimke kao rastera u programski jezik R može se izvesti pomoću više naredbi od kojih su neke *rast(), raster()* i *stack()* . U ovom radu korištena je funkcija *stack()* kojom se dobiva *RasterStack* objekt (*RasterStack* je zbirka *RasterLayer* objekata s istim prostornim opsegom (eng. *extent)* i razlučivosti) koji u ovom slučaju ima 6 slojeva jer koristimo 6 pojaseva satelitske snimke. Način na koji je učitana satelitska kompozitna snimka od 6 pojaseva prikazan je u sljedećem kodu (s obzirom na ograničeni prostor zbog duže putanje do *file-ova* u ovom je radu putanja označena sa *'filepath')*:

```
sentinel2 <- stack("'filepath'/gradzagrebHTRS.tif")
```

Točkasti trening podaci u okviru nadzirane klasifikacije učitani su funkcijom *readOGR()* koja je dostupna u okviru *terra* paketa, ali i nekih drugih paketa unutar R softvera. Učitavanje *shp* file-a sa trening podacima prikazano je u sljedećem kodu:

```
features <- readOGR("'filepath'","trening-tocke500HTRS")
```

7.4 Nenadzirana klasifikacija k-means algoritmom

Nenadzirana klasifikacija pomoću K-means algoritma izvedena je u R-studio-u. Korišten je NDVI indeks kao osnova za klasifikaciju. Izračunate su vrijednosti NDVI-a pomoću formule (3-2) za dati raster te se iz vrijednosti NDVI-a provela daljnja nenadzirana klasifikacija. Uvedena je nova varijabla "nr" koja preuzima vrijednosti NDVI-a. Korišten je Lyod algoritam koji je jedan od dostupnih algoritama unutar K-means funkcije.

Potrebno je napomenti kako je zbog preglednosti i općenite estetike ovog diplomskog rada u opisu svih metoda prikazan samo dio koda koji je određen kao najbitniji za shvaćanje materije.

Imenovanje pojaseva i izračun NDVI indeksa može se dobiti upotrebom slijedećeg koda:

```
names(sentinel2) <- c('blue', 'green', 'red', 'NIR', 'SWIR1',
'SWIR2')
ndvi<-(sentinel2[['NIR']]-
sentinel2[['red']])/(sentinel2[['NIR']] + sentinel2[['red']])
nr <- values(ndvi)
```

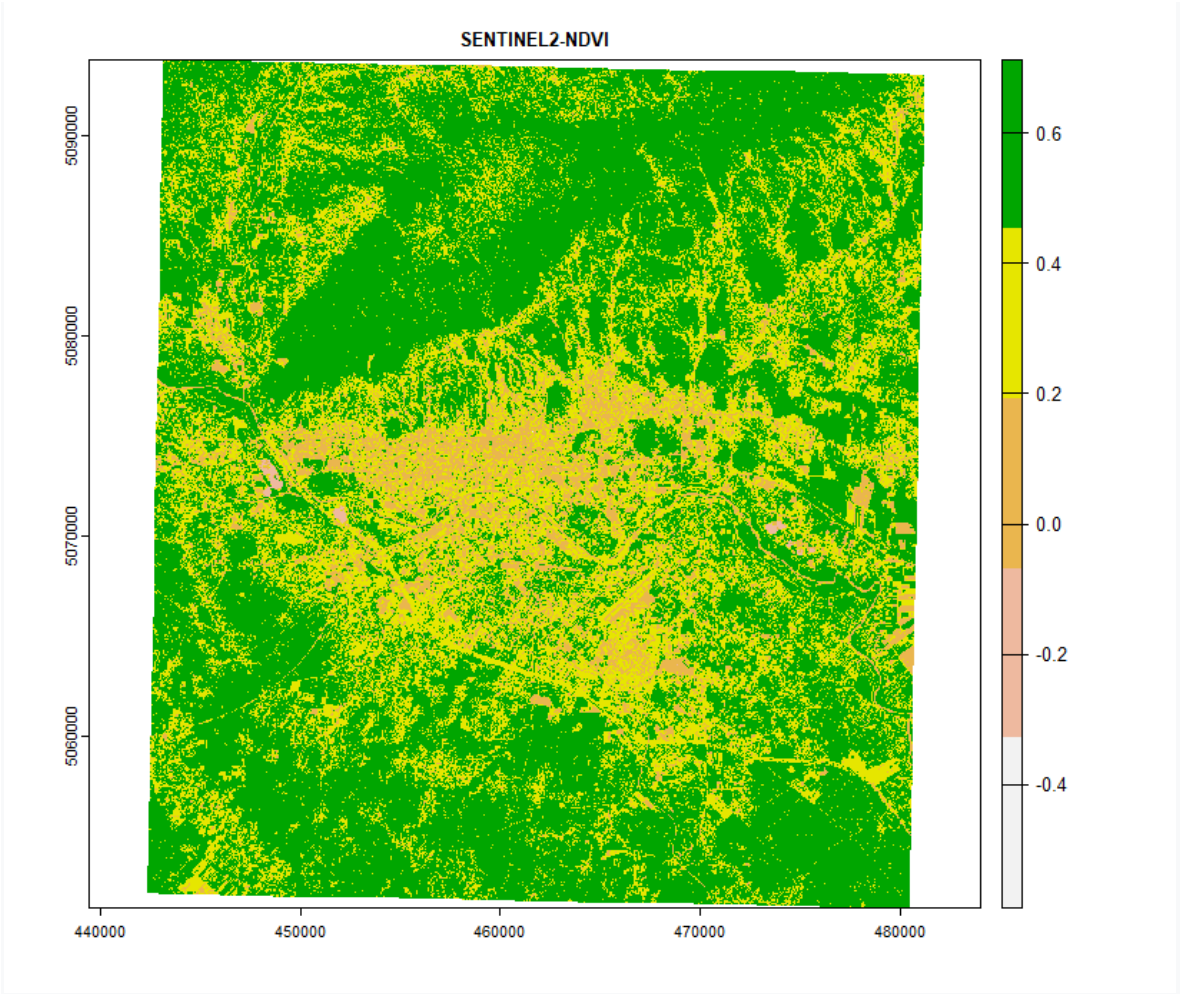
Sama funkcija *K-means* poziva se pomoću sljedećeg koda:

```
kmncluster <- kmeans(na.omit(nr), centers = 5, iter.max =
500, nstart = 2, algorithm="Lloyd")
```

Gdje se funkcija na.omit koristi kako bi se izbacile vrijednosti koje nedostaju (eng. *NA value*), *centers* označava broj centroida koje želimo te je odabrano 5 s obzirom na željeni broj klasa. Odabran je maksimalni broj iteracija 500 i Lloyd algoritam.

NDVI prikaz dobiva se pomoću funkcije *plot()*, i prikazan je na slici (Slika 7.6):

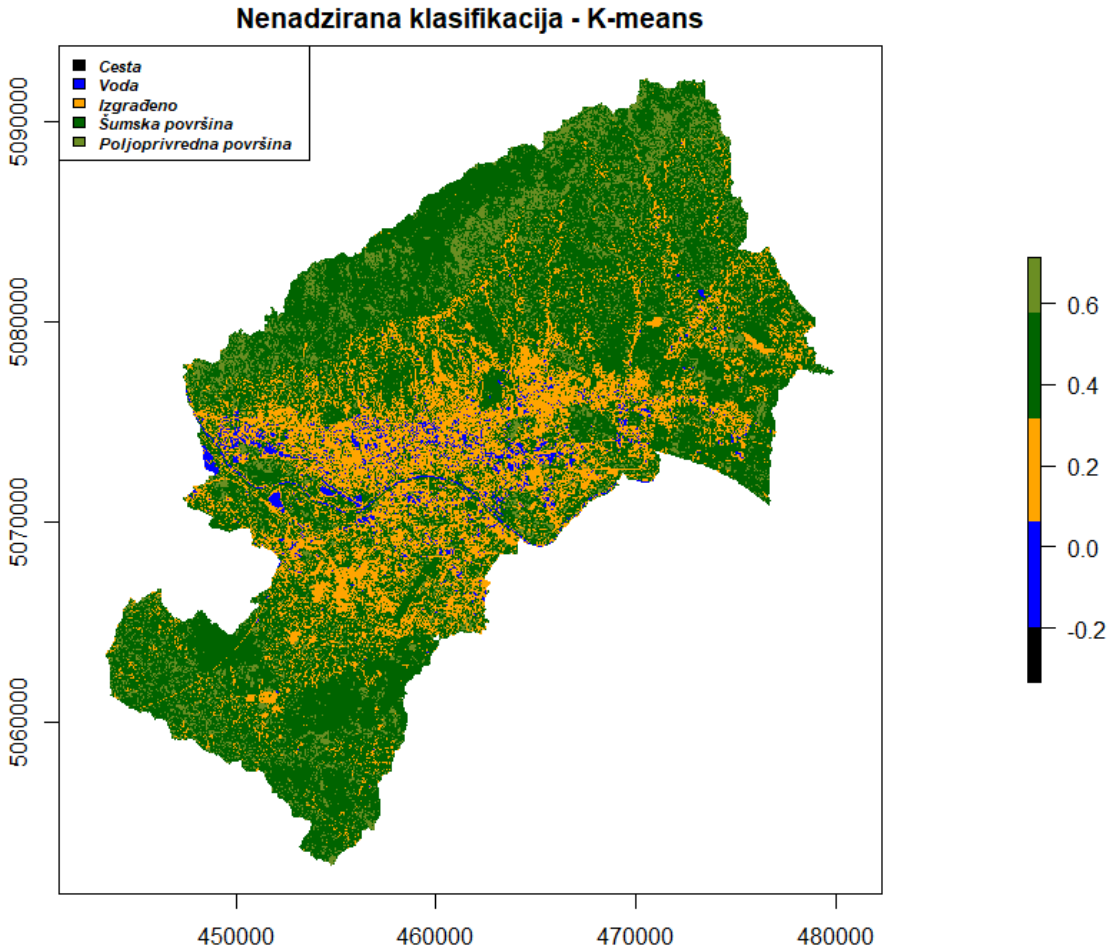
```
  plot(ndvi, col=rev( terrain.colors(5)), main="SENTINEL2-
NDVI")
```



Slika 7.6 *NDVI prikaz područja istraživanja*

Grafički prikaz rezultata k-means algoritma dobili smo koristeći funkciju *plot()* u programskom jeziku R i prikazan je na slici (Slika 7.7). Uz to je područje prilikom *plotanja* izrezano na granicu Grada Zagreba za što je korištena funkcija *mask()* i *crop_extent()* kao što je prikazano u sljedećem kodu:

```
crop_extent <-readOGR("'filepath'/GranicaZagrebaVector.shx")
masked <- mask(x = result, mask = crop_extent )
plot(masked,col=c("orange","black","darkgreen","olivedrab","blue"))
```



Slika 7.7 *Rezultat nenadzirane klasifikacije K-means*

Na grafičkom prikazu klasifikacije vidljivo je kako je K-means algoritam uspio odrediti određene klase, ali dolazi do značajnog miješanja nekih klasa (izgrađena površina i voda) dok neke praktički nije ni prepoznao (cesta). To nije neočekivano s obzirom da je K-means u svojoj osnovi klastering, a ne klasifikacijski algoritam.

7.5 Nadzirana klasifikacija random forest algoritmom

Ulazni podaci za klasifikaciju su Sentinel-2 satelitska snimka koja je učitana kao „data1“ i točkasti trening podaci u shp formatu učitani kao „features“.

Učitani podaci i njihove karakteristike prikazane su na slici (Slika 7.8)

```
> data1
class      : RasterStack
dimensions : 4321, 3807, 16450047, 6  (nrow, ncol, ncell, nlayers)
resolution : 10.18783, 9.813943  (x, y)
extent     : 442334.3, 481119.3, 5051429, 5093835  (xmin, xmax, ymin, ymax)
crs        : +proj=tmerc +lat_0=0 +lon_0=16.5 +k=0.9999 +x_0=500000 +y_0=0 +ellps=GRS80 +units=m +no_defs
names      : layer.1, layer.2, layer.3, layer.4, layer.5, layer.6
min values :      91,     500,     724,    1104,    1054,     992
max values :   18256,   17472,   16976,   16433,   16081,   16037

> features
class       : SpatialPointsDataFrame
features    : 500
extent      : 444970, 472480.4, 5065549, 5093577  (xmin, xmax, ymin, ymax)
crs         : +proj=tmerc +lat_0=0 +lon_0=16.5 +k=0.9999 +x_0=500000 +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
variables   : 1
names       : id
min values  :  1
max values  :  5
>
```

Slika 7.8 *Karakteristike učitanih varijabli za model random forest*

Vidljivo je da je *data1 "RasterStack"* objekt dok je *features* je skup prostornih točkastih podataka i sadrži točno 500 podatka podijeljenih u 5 klasa.

Sljedeći korak je izvući vrijednosti piksela na lokaciji trening podataka te rezultate zapisati u obliku *csv* file-a. To je učinjeno naredbom *ExtractByPoint()* koja je prisutna unutar paketa "ExtractTrainData". Zapisivanje vrijednosti u *csv* file napravljeno je jednostavnom naredbom *write.csv()*.

Opisana radnja prikazana je u kodu:

```
Out.colName<-In.colName<-"id"

e <- ExtractByPoint(data1,features,In.colName,Out.colName)

e

write.csv(e,"'filepath' /trening_input500.csv")
```

Kao idući korak potrebno je učitati trening podatke za koje je izračunata vrijednosti piksela te je to učinjeno slijedećim kodom:

```
training <- read.csv("'filepath'/trening_input500.csv")
training
```

Nakon toga odredio se postotak podataka koji će bit korišten za trening i testiranje. S obzirom na uobičajenu praksu u području strojnog učenja odabrana je vrijednost 70% za trening podatke te vrijednost 30% za test podatke. Opisani postupak izveden je kodom:

```
set.seed(190)
id <- sample(2, nrow(d), replace=TRUE, prob=c(0.7,0.3))
train <-d[id==1,]
test<-d[id==2,]
train
test
```

S obzirom da su sada trening podaci spremni za ulazak u model moguće je krenuti u izgradnju Random Forest modela. Na slijedećem kodu prikazan je model koji je korišten za klasifikaciju:

```
#RF Model
CV <- trainControl(method ="cv",
                   number=10,
                   savePredictions=TRUE
                   )

rfGrid <- expand.grid(mtry=(1:10))
rf <- train(id~.,data=train,
            method="rf",
            trControl=CV,
            verbose=FALSE,
            tuneGrid=rfGrid,
            importance=TRUE)

rf
plot(rf)
varImp(rf, scale=FALSE)
```

Metoda za kontroliranje trening procesa je *„cv“* odnosno kros-validacija. To znači da se za trening koriste podskupovi trening podataka te se procjenjuju na temelju komplementarnih podskupova. *rfGrid* definira takozvani *tuning grid* koji se odnosi na proces prilagodbe hiperparametara modela kako bi se poboljšala njegova izvedba na danom skupu podataka. Stvorena je mreža različitih kombinacija hiperparametara, a metoda kros-validacije koristi se za procjenu izvedbe modela za svaku kombinaciju. Hiperparametri u *random forest* algoritmima uključuju broj *decision tree-a,* njihovu maksimalnu dubinu i minimalni broj uzoraka potrebnih za dijeljenje čvora. (*Towardsdatascience-a, 2020*.). Potom je model podvrgnut treningu naredbom *train().*

Sama predikcija pomoću izgrađenog modela dobivena je funkcijom *predict()* što je prikazano u kodu:

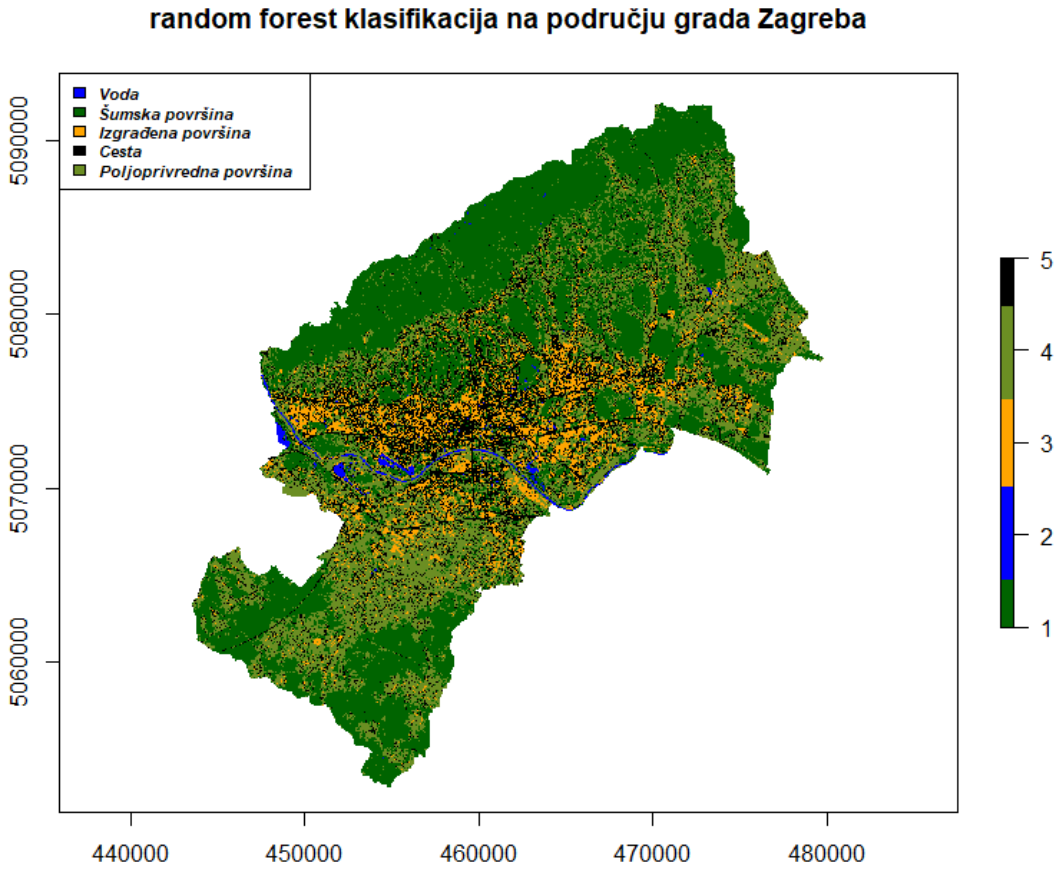
```
classified <- predict(data1,rf, type='raw',progress='window')
```

Grafički prikaz nadzirane klasifikacije dobiven je naredbom *plot()* dok je legenda generirana naredbom *legend().* Također, kao i u prethodnom slučaju, prikaz je izrezan na područje Grada Zagreba. Opisani postupak prikazan je u sljedećem kodu:

```
crop_extent <-
readOGR("C:/Users/NinjaBoi/Documents/GranicaZagrebaVector.shx
")
masked1 <- mask(x = classified, mask = crop_extent )
plot(masked1, main="Nadzirana klasifikacija-
randomForest",col=c("darkgreen", "blue",
"orange","olivedrab","black"))
legend("topleft", legend=c("Voda", "Šumska površina",
"Izgradeno","Cesta","Poljoprivredna površina"),
```

```
        fill=c("blue",
"darkgreen","orange","black","olivedrab"), text.font=4,
bg='white',cex=0.7)
```

Navedeni kod daje grafičko rješenje prikazano na slici (Slika 7.9).



Slika 7.9 *Rezultat random forest klasifikacije*

S obzirom na grafički prikaz klasifikacije jasno je da nadzirani model random forest daje znatno bolje rezultate od prethodnog nenadziranog modela. Jasno su izdvojene ceste, vodena i šumska površina te čak poljoprivredne i izgrađene površine.

7.6 Nadzirana klasifikacija xgboost algoritmom

Početni podaci učitani su na isti način kao i za *random forest* klasifikaciju sa razlikom u nazivu varijabli gdje je *ras* satelitska snimka, a *shp* predstavlja točkaste trening podatke u *shp* obliku. Podaci su dakle u potpunosti jednaki osim što su drugačije nazvane varijable.

Karakteristike navedenih varijabli prikazane su na slici (Slika 7.10).

```
> ras
class      : RasterStack
dimensions : 4321, 3807, 16450047, 6  (nrow, ncol, ncell, nlayers)
resolution : 10.18783, 9.813943  (x, y)
extent     : 442334.3, 481119.3, 5051429, 5093835  (xmin, xmax, ymin, ymax)
crs        : +proj=tmerc +lat_0=0 +lon_0=16.5 +k=0.9999 +x_0=500000 +y_0=0 +ellps=GRS80 +units=m +no_defs
names      : gradzagrebHTRS_1, gradzagrebHTRS_2, gradzagrebHTRS_3, gradzagrebHTRS_4, gradzagrebHTRS_5, gradzagrebHTRS_6

> shp
class       : SpatialPointsDataFrame
features    : 500
extent      : 444970, 472480.4, 5065549, 5093577  (xmin, xmax, ymin, ymax)
crs         : +proj=tmerc +lat_0=0 +lon_0=16.5 +k=0.9999 +x_0=500000 +y_0=0 +ellps=GRS80 +units=m +no_defs
variables   : 1
names       : id
min values  :  1
max values  :  5
>
```

Slika 7.10 *Karakteristke učitanih varijabli za model xgboost*

Vrijednosti piksela izvučene su istom naredbom kao i kod *random forest* algoritma te su stvoreni trening podaci u obliku matrice:

```
Out.colName<-In.colName<-"id"
vals <- ExtractByPoint(ras,shp,In.colName,Out.colName)
train <- data.matrix(vals)
```

Podjela trening podataka na trening (70% ukupnog broja trening podataka) i test (30% ukupnog broja trening podataka) napravila se pomoću naredbe *createDataPartition()* u okviru paketa *„caret“* kodom:

```
set.seed(123)
train_ind <- createDataPartition(classes, p = 0.7, list =
FALSE)
train_data <- train[train_ind, ]
train_classes <- classes[train_ind]
test_data <- train[-train_ind, ]
```

```
test_classes <- classes[-train_ind]
```

Sam *xgboost* algoritam dostupan je kao paket *xgboost* u R-u stoga je potrebno samo definirati parametre:

```
xgb <- xgboost(data = train_data,
               label = train_classes,
               eta = 0.1,
               max_depth = 6,
               nround=100,
               objective = "multi:softmax",
               num_class = length(unique(classes)),
               nthread = 5)
```

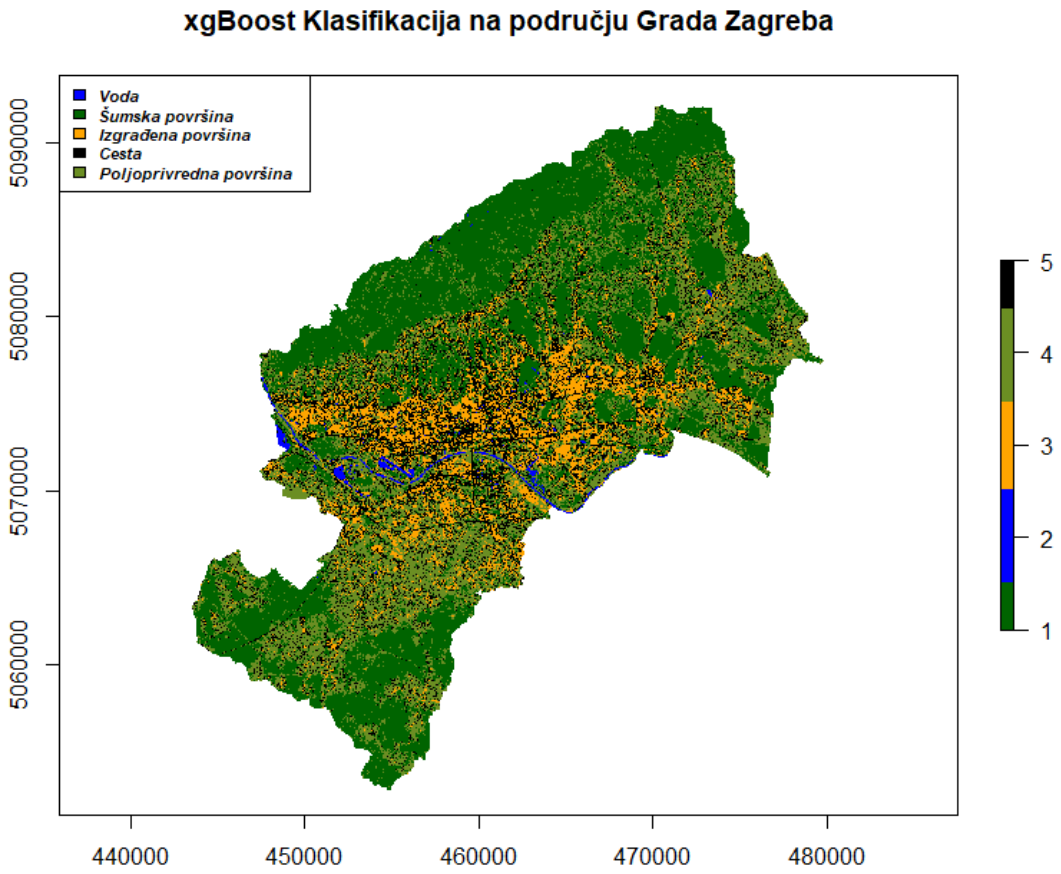
Predikcija nepoznatih vrijednosti i klasifikacija dobivena je sljedećim kodom:

```
result   <- predict(xgb, test_data)
res    <- raster(ras)
result   <- predict(xgb, ras[1:(nrow(ras)*ncol(ras))])
res   <- setValues(res,result+1)
```

Grafički prikaz klasifikacije prikazan je naredbom *plot()* te je izrezan na područje Grada Zagreba:

```
crop_extent <- readOGR("'filepath'/GranicaZagrebaVector.shx")
masked <- mask(x = res, mask = crop_extent )
myColor <- c("darkgreen","blue","orange","olivedrab","black")
plot(masked, main="xgBoost klasifikacija", col=myColor)
legend("topleft", legend=c("Voda", "Šumska površina", "Izgrađena
površina","Cesta","Poljoprivredna površina"),
fill=c("blue", "darkgreen","orange","black","olivedrab"), text.font=4,
bg='white',cex=0.7)
```

Grafički prikaz klasifikacije prikazan je na slici (Slika 7.11).

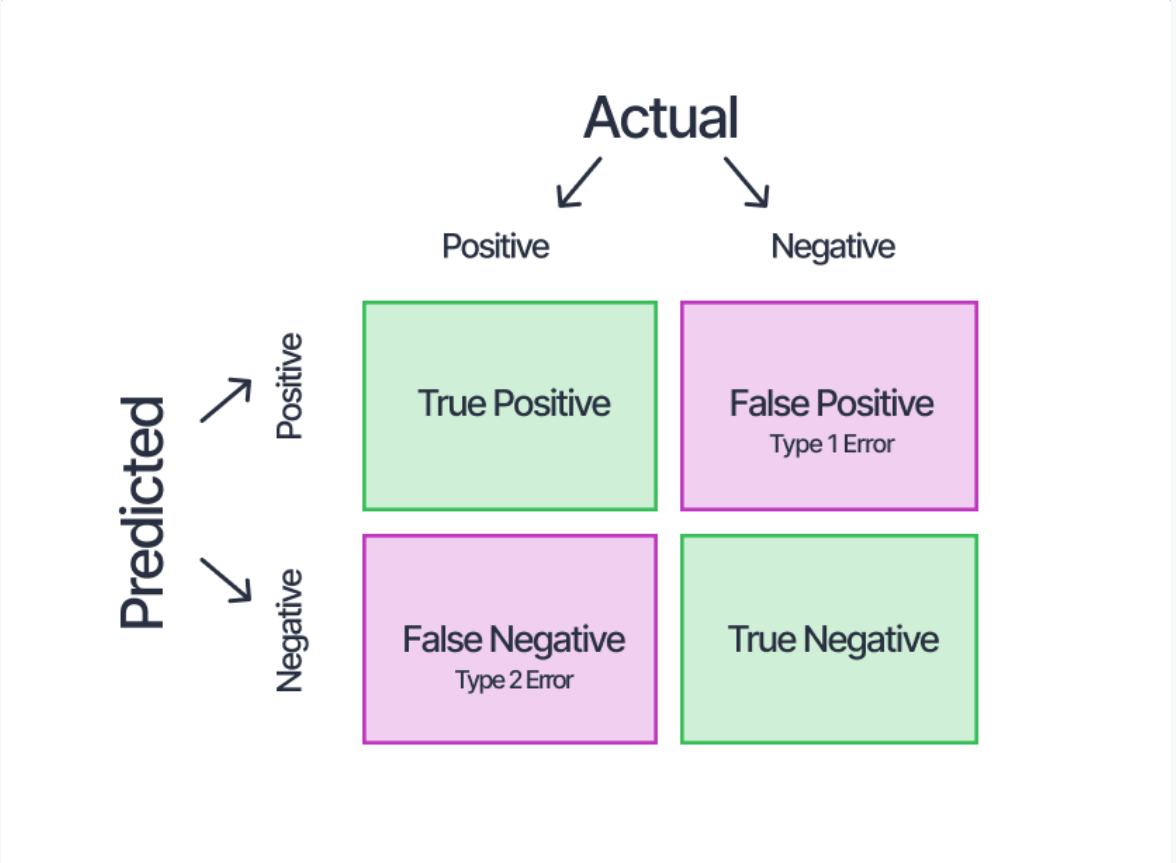


Slika 7.11 *Rezultat xgboost klasifikacije*

Vizualnim pregledom ne može se jasno raspoznati razlika u točnosti između rezultata *xgboost i random forest*. Međutim, svakako je jasno da oba dvije nadzirane metode daju preciznije rezultate od nenadzirane *K-means* metode.

8. USPOREDBA REZULTATA NADZIRANE KLASIFIKACIJE

S obzirom da je nenadzirana klasifikacija pokazala značajnu razinu nepreciznosti što je vidljivo u grafičkom prikazu na slici (Slika 7.7) provedena je statistička usporedba samo između dvije nadzirane metode *random forest* i *xgboost*. Kako su točkasti trening podaci prije provedbe samih algoritama bili podijeljeni u trening i test skupinu s omjerom 70 naprema 30 moguće je procijeniti preciznost algoritama uspoređivanjem predviđenih vrijednosti sa referentnim (test) vrijednostima. Takvu usporedbu moguće je izvesti takozvanom matricom zabune (eng. *confusion matrix*). *Confusion matrix* je mjera izvedbe za klasifikaciju strojnog učenja u problemima gdje rješenje može biti dvije ili više klasa. (*Towardsdatascience-b, 2018.*). Shematski prikaz *confusion matrix-a* sa dvije klase prikazan je na slici (Slika 8.1).



Slika 8.1 *Shematski prikaz confusion matrix-a sa dvije klase (V7labs, 2023.,)*

Unutar programskog jezika R, *confusion matrix* moguće je stvoriti naredbom *confusionMatrix(„predviđene klase“, „referentne klase“)* koja je dostupna u okviru paketa *„caret“*.

Confusion matrix kod *random forest* modela stvorilo se sljedećim kodom:

```
predicted_class_test <- predict(rf,test)
predicted_class_test
test[,7]
confusionMatrix(predicted_class_test,test[,7])
```

Confusion matrix xgboost modela dobivena je kodom:

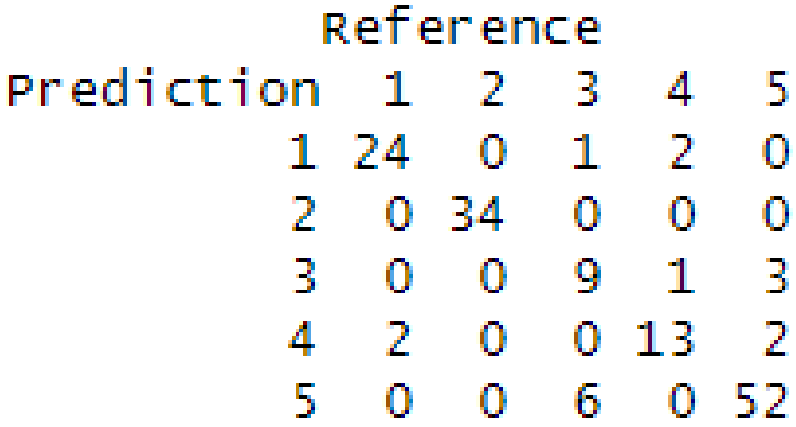
```
result   <- predict(xgb, test_data)
result   <- as.factor(result + 1)
test_classes <- as.factor(test_classes + 1)


confusionMatrix(result, test_classes)
```

Rezultati i sam izgled *confusion matrix-a* u R konzoli za obje klasifikacije vidljivi su na slikama (Slika 8.2) i (Slika 8.3)

```
          Reference
Prediction  1  2  3  4  5
         1 32  0  1  3  0
         2  0 39  0  0  0
         3  0  0  5  2  0
         4  3  0  0 12  1
         5  1  0  5  0 55
```

Slika 8.2 *Confusion matrix za random forest klasifikaciju*



Slika 8.3 *Confusion matrix za xgboost klasifikaciju*

Uz samu matricu, u okviru naredbe *confusionMatrix()* program stvara i odjeljak „statistika po klasi“ te prikazuje sljedeće veličine:

- Osjetljivost (poznata i kao *recall*): udio pozitivnih slučajeva koji su ispravno identificirani kao pozitivni;
- Specifičnost (također poznata kao *True Negative Rate*): udio negativnih slučajeva koji su ispravno identificirani kao negativni;
- Pozitivna prediktivna vrijednost (također poznata kao preciznost): udio pozitivnih predviđanja koja su točna;
- Negativna prediktivna vrijednost: udio negativnih predviđanja koja su točna;
- Prevalencija: udio pozitivnih slučajeva u ukupnim podacima;
- Stopa otkrivanja: udio pozitivnih slučajeva koji su otkriveni;
- Prevalencija otkrivanja: udio pozitivnih slučajeva koji su otkriveni ili predviđeni kao pozitivni.

Uravnotežena točnost: prosječna točnost uzimajući u obzir obje klase kao jednako važne.

Ove statistike pružaju pregled izvedbe klasifikatora za svaku klasu. Rezultati za obje klasifikacije prikazane su na slikama (Slika 8.4) i (Slika 8.5)

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.8889	1.0000	0.45455	0.70588	0.9821
Specificity	0.9675	1.0000	0.98649	0.97183	0.9417
Pos Pred Value	0.8889	1.0000	0.71429	0.75000	0.9016
Neg Pred Value	0.9675	1.0000	0.96053	0.96503	0.9898
Prevalence	0.2264	0.2453	0.06918	0.10692	0.3522
Detection Rate	0.2013	0.2453	0.03145	0.07547	0.3459
Detection Prevalence	0.2264	0.2453	0.04403	0.10063	0.3836
Balanced Accuracy	0.9282	1.0000	0.72052	0.83886	0.9619

Slika 8.4 *Statistika po klasi za random forest klasifikaciju*

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.9231	1.0000	0.56250	0.81250	0.9123
Specificity	0.9756	1.0000	0.96992	0.96992	0.9348
Pos Pred Value	0.8889	1.0000	0.69231	0.76471	0.8966
Neg Pred Value	0.9836	1.0000	0.94853	0.97727	0.9451
Prevalence	0.1745	0.2282	0.10738	0.10738	0.3826
Detection Rate	0.1611	0.2282	0.06040	0.08725	0.3490
Detection Prevalence	0.1812	0.2282	0.08725	0.11409	0.3893
Balanced Accuracy	0.9493	1.0000	0.76621	0.89121	0.9235

Slika 8.5 *Statistika po klasi za xgboost klasifikaciju*

Iz prikazanih statističkih procijena izvedbe modela vidljivo je da su obje metode izuzetno precizne te su razlike u točnosti minimalne. Uz to vidi se da je preciznost po klasama također vrlo slična odnosno klasa koju je svaki model najtočnije odredio je voda i iznosi 100% koju potom po točnosti prate klasa 1 odnosno šumska površina, klasa 5 odnosno cesta te klasa 4 koju čine poljoprivredne površine i klasa 3 odnosno izgrađena površina sa najmanjom točnosti predikcije.

Kako bi se bolje protumačili rezultati pregledalo se koliko je trening podataka bilo pridruže no svakoj klasi te se napravio spektralni profil (Slika 8.5) kako bi se vidjela preklapanja u r efleksiji određenih klasa.

Tablica 8-1 *Početna količina trening podataka po klasi*

1 – Šumska površina	2 - voda	3 – izgrađena površina	4 – poljoprivredna površina	5 - cesta
93	108	48	60	191

S obzirom na podatke iz tablice (Tablica 8-1) i slika (Slika 8.3), (Slika 8.4) možemo vidjeti da postoji povezanost između količine trening podataka i preciznosti klasifikacijskog algoritma. Za klasu sa najmanje trening podataka odnosno njih 48 (klasa 3 – izgrađena površina) klasifikacijski algoritam dobio je najmanju točnost predikcije. Dok je za klasu voda koja sadrži 108 trening podataka dobio najveću razinu točnosti od 100%.

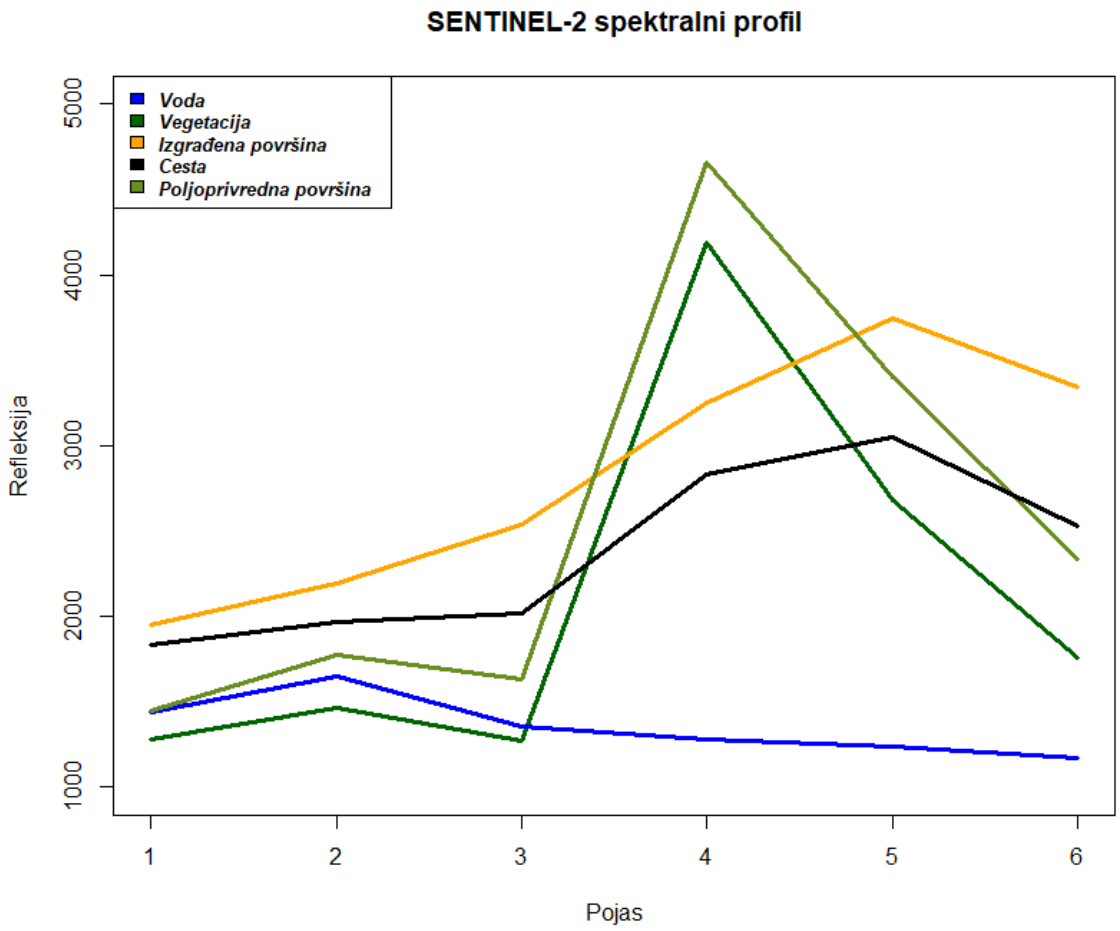
Spektralni profil na slici (Slika 8.6) dobiven je pomoću sljedećeg koda u programskom jeziku R:

```
shp$id <- over(shp, shp)$id
head(vals)
ms <- aggregate(vals, list(shp$id), mean)
rownames(ms) <- ms[,1]
ms <- ms[,-1]
ms
mycolor <- c("darkgreen","blue","orange","olivedrab","black")

#transformacija ms iz data frame-a u matrix
ms <- as.matrix(ms)

# stvori prazan plot
plot(0, ylim=c(1000,5000), xlim = c(1,6), type='n', xlab="Band", ylab =
"Refleksija")

# dodaj klase
for (i in 1:nrow(ms)){
  lines(ms[i,], type = "l", lwd = 3, lty = 1, col = mycolor[i])
}
# Naslov
title(main="SENTINEL-2 spektralni profil", font.main = 2)
# Legenda
legend("topleft",legend=c("Voda","Vegetacija","Izgradeno","Cesta","Poljop
rivrednapovršina"),fill=c("blue","darkgreen","orange","black","olivedrab"
), text.font=4, bg='white',cex=0.8)
```



Slika 8.6 *Prikaz SENTINEL-2 spektralnog profila za 6 pojaseva*

Na grafičkom prikazu spektralnog profila vidljiva je refleksija svake klase s obzirom na određeni pojas. Vidljivo je da se profili dosta dobro razlikuju, ali su jasna neka preklapanja te je očito da su oblici krivulje za klase izgrađenog područja i cesta sličnog izgleda. Isto vrijedi za šumske i poljoprivredne površine što ukazuje na to da bi se te klase mogle teže izdvojiti jedna od druge. Uz to, može se primjetiti je da refleksija za klasu vode prati zaseban "trend" te je jasno izdvojena od drugih, pogotovo u pojasevima 4, 5 i 6. To bi mogao biti razlog zbog čega je, uz veći broj trening podataka, ta klasa predviđena sa 100%-tnom točnošću dok druge klase nisu bile tako precizno predviđene.

U okviru *confusionMatrix()* naredbe u R-u dobiju se i dvije potencijalno korisne veličine za determinaciju točnosti predikcije modela: preciznost i takozvana "kappa" vrijednost. Dok je preciznost mjera točnosti predikcije modela u odnosu na stvarne vrijednosti, kappa uzima u obzir distribuciju klasa u podacima i daje mjeru slaganja između predviđenih i stvarnih oznaka klasa. Ona se kreće od -1 do 1, gdje 1 označava savršeno slaganje, 0 označava slučajno slaganje, a negativne vrijednosti označavaju manje slaganje nego što se slučajno očekivalo. S obzirom da naš skup trening podataka nije u potpunosti jednak za sve klase ona nam je korisna za bolju interpretaciju rezultata.

Vrijednosti preciznosti i kappa za svaku metodu prikazana je u tablici (Tablica 8-2).

Tablica 8-2 *Prikaz preciznosti i kappa vrijednosti za svaki model*

	Preciznost	Kappa
random forest klasifikacija	0.8994	0.8639
xgboost klasifikacija	0.8859	0.8470

Vidljivo je da razlika preciznosti iznosi 0.0135 to jest 1.35% u korist random forest klasifikacije. Također može se izračunati razlika kappa vrijednosti koja iznosi 0.0169 odnosno 1.69% također u korist random forest klasifikacije.

9. ZAKLJUČAK

Provedbom nenadzirane i nadzirane klasifikacije SENTINEL-2 satelitske snimke na području Grada Zagreba jasno je vidljivo, već iz vizualnog prikaza, da su nadzirane metode značajno preciznije u raspoznavanju i predikciji određenih klasa. To je i očekivano s obzirom da se za takve algoritme ipak koriste trening podaci o čijoj kvaliteti i preciznosti također ovisi i izvedba te krajnji rezultat samog modela. Izgradnja nadziranog modela klasifikacije je zahtjevnija s obzirom da uključuje stvaranje i korištenje trening podataka, različite hiperparametre i duže vrijeme koje potrebno računalu da izvede klasifikaciju zbog kompleksnosti algoritma. Nenadzirana klasifikacija dala je rješenje koje pokazuje klase sa određenom razinom preciznosti (šumska površina, voda, izgrađena površina) dok druge (posebno ceste, ali i poljoprivredne površine) praktički uopće nije ili je slabo razlučila. To nije neočekivano s obzirom da je K-means algoritam u svojoj osnovi klastering, a ne klasifikacijski algoritam, iako se može kao u primjeru ovog rada koristiti za nenadziranu klasifikaciju. U slučaju nadzirane klasifikacije, preciznost dobivena procjenom točnosti iznosi gotovo 90% odnosno 88.59% za xgboost i 89.94% za random forest klasifikaciju. Iz toga, određeno je da je razlika između preciznosti predikcije random forest modela i xgboost modela je 0.0135 odnosno 1.35%. Tumačenju rezultata analize točnosti treba pristupiti s oprezom zbog ovisnosti modela o kvaliteti trening podataka. Potrebno je uzeti u obzir da na preciznost modela utječu kako trening podaci, tako i test podatci koji se uzimaju kao referentne vrijednosti. Generalno govoreći, veći skup trening podataka vjerojatno bi rezultirao još većom razinom točnosti klasifikacijskih modela. Uz ukupan broj trening podataka važna je i raspodjela podatka po klasama. Parametri poput vrijednosti „kappa“ uzimaju u obzir distribuciju trening podataka po klasama, ali bi ujednačeni raspored doveo do veće pouzdanosti klasifikacijskog modela. S obzirom na to, može se preporučiti da se koristi približno jednak broj trening podataka za svaku klasu. Uzimajući u obzir broj korištenih trening podataka (500) u ovom radu preciznost koja premašuje 85% u oba modela je zadovoljavajuća te se random forest model pokazao nešto preciznijim u ovom specifičnom slučaju. Dodatnim podešavanjem hiperparametara teoretski bi se mogla dobiti i veća preciznost koristeći isti skup trening podataka te se oba modela mogu preporučiti za nadziranu klasifikaciju satelitskih snimki.

10. LITERATURA

AI Pool, 2021., *Random Forests understanding.* URL: https://ai-pool.com/a/s/random-forests-understanding (1.2.2023.)

BMC, 2021. , *Bias & Variance in Machine Learning: Concepts & Tutorials* URL: https://www.bmc.com/blogs/bias-variance-machine-learning/ (28.1.2023.)

Brainkart, 2023., *Elements of remote sensing. URL: https://www.brainkart.com/article/Elements-of-Remote-Sensing_41124/* (28.1.2023.)

Copernicus-a, 2023., *Sentinel User Guides.* URL: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/overview (10.2.2023.)

Copernicus-b, 2023., *MultiSpectral Instrument (MSI) overview.* URL: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/msi-instrument (10.2.2023.)

Custom-scripts, 2023., *Normalized difference vegetation index.* URL: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndvi/ (10.2.2023.)

Datacamp**,** 2019., *R Packages: A Beginner's Tutorial.* URL: https://www.datacamp.com/tutorial/r-packages-guide (30.1.2023.)

Datascience lab, 2013., *Clustering with K-means in Python.* URL: https://datasciencelab.wordpress.com/tag/lloyds-algorithm/ (28.1.2023.)

Displayr, 2023., *Gradient boosting explained.* URL: https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/https (2.2.2023.)

ESA, 2023., *Sentinel-2.*URL: https://sentinel.esa.int/web/sentinel/missions/sentinel-2 (31.1.2023.)

GIS Lounge, 2022., *What is GIS?* URL: https://www.gislounge.com/what-is-gis/ (17.1.2023.)

Halapir, I., 2022. *Analiza podložnosti na klizanje primjenom bivarijantnih statističkih metoda na području podsljemenske zone Grada Zagreba.* Diplomski rad. Zagreb: Sveučilište u Zagrebu, Rudarsko-geološko-naftni fakultet.

IBM, 2023., *What is a decision tree?* URL: https://www.ibm.com/topics/decision-trees (29.1.2023.)

Iowa State University, 2014., *Introduction to QGIS.* URL: https://store.extension.iastate.edu/product/Introduction-to-QGIS (10.2.2023.)

Javapoint-a, 2023., *R-programming tutorial.* URL: https://www.javatpoint.com/r-tutorial (1.2.2023.)

Javapoint-b, 2023., *K-means clustering algorithm.* URL: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning (1.2.2023.)

Javapoint-c, 2023., *Decision tree classification algorithm.* URL: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (10.2.2023.)

Levizzani, V. i dr., 2019., *Satellite remote sensing of precipitation and the terrestrial water cycle in a changing climate.* Remote sensing, *11*(19), 2301. URL: https://www.mdpi.com/2072-4292/11/19/2301 (25.1.2023.)

Loures, L. i dr., 2020., *Assessing the effectiveness of precision agriculture management systems in mediterranean small farms*. Sustainability, *12*(9), 3765. URL: https://www.researchgate.net/figure/Example-of-the-use-of-a-normalized-difference-vegetation-index-NDVI_fig1_341188077

NASA-a, 2014., *What is a satellite? URL: https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-satellite-58.html* (28.1.2023.)

NASA-b, 2020., *What is a the Internation Space Station? URL: https://www.nasa.gov/audience/forstudents/k-4/stories/nasa-knows/what-is-the-iss-k4.html* (10.2.2023.)

RSpatial, 2023., *Remote sensing with terra*. URL: https://rspatial.org/rs/index.html (24.1.2023.)

RStudio, 2023., *Download the R Studio IDE, URL: https://support--rstudio-com.netlify.app/products/rstudio/download/* (26.11.2023.)

Science direct, 2021., *Extreme gradient boosting model to predict groundwater levels in Selangor Malaysia.* URL: https://www.sciencedirect.com/science/article/pii/S2090447921000125 (29.1.2023.)

Stackexchange, 2014., *What is the intuition behind kappa statistical value? URL: https://stats.stackexchange.com/questions/124001/what-is-the-intuition-behind-the-kappa-statistical-value-in-classification* (1.2.2023.)

Towardsdatascience-a, 2020., *Parameters and hyperparameters in machine learning and deep learning*. URL: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (1.2.2023.)

Towardsdatascience-b, 2018., *Understanding confusion matrix*. URL: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (1.2.2023.)

Tutorialspoint, 2023., *R tutorial*. URL: https://www.tutorialspoint.com/r/index.htm (13.1.2023.)

Unearth labs, 2023., *What is QGIS*. URL: https://www.unearthlabs.com/ (28.1.2023.)

USGS, 2023., *What is remote sensing and what is it used for?* URL: https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used (10.2.2023.)

V7labs, 2023., *Confusion Matrix: How to use it & Interpret results.* URL: https://www.v7labs.com/blog/confusion-matrix-guide (10.2.2023.)